Mental health detection via social media

Dhanush R
Department of Artificial intelligence and Data science
R.M.K Engineering College
Chennai, India
rdha21228.ad@rmkec.ac.in

Pavan Sai Reddy
Department of Artificial intelligence and Data science
R.M.K Engineering College
Chennai, India
eswale221801.ad@rmkec.ac.in

Gnana Muhelan
Department of Artificial intelligence and Data science
R.M.K Engineering College
Chennai, India
gnanle221804.ad@rmkec.ac.in

Mr. K. T. Dhanasekaran
Department of Artificial intelligence and Data science
R.M.K Engineering College
Chennai, India
ktd.ad@rmkec.ac.in

***Abstract*— The increasing prevalence of mental health issues, especially depression and anxiety, has underscored the need for accessible and scalable mental health support systems. Social media and messaging platforms serve as popular means of communication, where individuals often express their emotions, including signs of mental distress. Maṉa is a web-based application designed to analyze mental health via social media interactions. The system leverages user-generated comments on posts and tweets to assess emotional well-being. By aggregating social media data through APIs or user uploads, Maṉa employs a fine-tuned RoBERTa model to perform sentiment analysis, classifying interactions as either positive or negative. When negative sentiments prevail, the system activates MaṉaNow—a dynamic questioning AI that conducts an in-depth mental health assessment through a series of targeted questions and ultimately generates a personalized report. In parallel, MaṉaChat provides an instructional conversational interface, powered by the meta llama/Llama-3.2-3B-Instruct model, to offer immediate stress-reduction strategies and mental health guidance. This dual approach not only enhances user engagement but also ensures timely intervention by coupling real-time analysis with supportive resources.**

Keywords -- RoBERTa, Emotions severity levels, chatbot, Mental health support, Social-media, Healthcare analytics, finetuning, Prompt tunning

INTRODUCTION

Presently the world wide mental health care system is going through challenging times. According to the World Health Organization, one in four people affected by mental illness at some point in their lives [1]. Mental disorder is still the leading cause of health-related economic hardship around the world [2]. In particular, depression and anxiety are the most frequent causes, affecting an estimated 322 million (depression) and 264 million (anxiety) individuals globally [3]. In spite of such growing burden, there seems to be an acute shortage of mental health professionals worldwide (9 per 100, 000 people), principally in Southeast Asia (2.5 per 100,000 people) [4]. Despite the fact that there are efficient and well-known therapies for numerous mental and neurological disorders, only half of people, afflicted by mental disorder, receive them [1]. The main obstacles to successful and wide ranging treatment have been highlighted as a lack of resources and qualified medical professionals, as well as social discrimination, stigma and marginalization [1]. Growing public expectations are raising the bar for healthcare systems to overcome the obstacles and offer an accessible, cost-effective, and evidence-based treatment to medically indigent individuals [5]. The synergy between ML and NLP allows us to harness the potential of this textual data fully. ML techniques, including text classification and sentiment analysis, enable automated categorization and sentiment assessment of patient notes, facilitating the extraction of valuable insights on patient satisfaction, concerns, and feedback. Simultaneously, NLP techniques such as topic modelling unveil latent themes and patterns within the text, providing a deeper understanding of the patient journey.

With the widespread use of social media, individuals increasingly share their thoughts, emotions, and moods online, creating a vast pool of data that can offer valuable insights into public sentiment. However, interpreting these emotions accurately is challenging due to the diverse linguistic expressions, informal language, and cultural variations that characterize social media posts. Traditional sentiment analysis methods often fall short, failing to capture the nuanced differences between similar emotions (e.g., sadness vs. disappointment) and the context-dependent nature of online interactions.

Large language models (LLMs) have been the focus of significant attention in the field of artificial intelligence (AI) in recent years. These models are trained on massive amounts of data and have demonstrated remarkable performance in natural language processing (NLP) tasks such as language generation, machine translation, and question-answering [6].

In mental health analysis, traditional methods mostly make predictions in a discriminative manner. Effective methods mostly finetune pre-trained language models (PLMs), such as

BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), on a small target set (Zhang et al., 2022; Ji, 2022) usually for one mental health condition. To further enhance the PLM representations, some works pre-train language models from scratch with large-scale mental health-related social media data, which usually produce better post representations than general PLMs. Representative works include MentalBERT (Ji et al., 2022b), MentalXLNet (Ji et al., 2023), etc.

Though the above black-box models achieve impressive classification performance, there are works exploring interpretable mental health analysis. Some works incorporate metaphor concept mappings as extra features to provide clues on model decisions (Han et al., 2022). Other works introduced PHQ-9 questionnaire information to assist the predictions (Nguyen et al., 2022; Zhang et al., 2023). Commonsense knowledge graphs were also leveraged to increase the transparency of PLMs (Harrigian et al., 2020; Yang et al., 2022). The recent advancements in LLMs take a leap forward for interpretable mental health analysis. Some works (Amin et al., 2023; Xu et al., 2023; Yang et al., 2023) comprehensively evaluated the performance of general foundation LLMs on various mental health analysis tasks. Xu et al. (Xu et al., 2023) glimpsed the explanation generation ability of LLMs, and Yang et al. (Yang et al., 2022) holistically evaluated ChatGPT's explanation generation ability with careful human evaluation.

I. LITERATURE SURVEY

We review several key areas that are closely related to the problem of emotion detection and mental health analysis.

Sentiment Analysis. Sentiment analysis aims at discovering the contextual polarity of the documents [Pang and Lee, 2008]. [Li et al., 2009] proposed a Non-negative Matrix Fac torization (NMF) approach which leverages lexical knowl edge for sentiment classification. Recent work [Bollen et al., 2011; Golder and Macy, 2011] has focused on mining temporal and seasonal trends of sentiment. Sentiment analy sis is a closely related problem, however emotions are much more expressive than sentiments. Moreover, emotions need not contain a sentiment and vice-versa [Liu, 2012]. Emotion Detection. Emotion models are primarily of two types [Ekkekakis, 2013]: (i) dimensional, and (ii) categori cal. Dimensional models represent emotions on three dimen sions: valence, arousal and dominance. [De Choudhury et al., 2012a] extended it to circumplex model and studied vari ous aspects of the relationship between mood expression and human behavior in social media. Categorical models repre sent emotions into finite categories. Ekman's basic emotion set (anger, disgust, fear, happiness, sadness, and surprise) is arguably the most popular emotion taxonomy [Ekman, 1992].

There was, however, one research work [7], that classified text into six Emotion Categories, but that was only limited to classification of news headlines, and the training set used was created manually. We, on the other hand, have developed a system, which classifies text in any form (eg. news, tweets, or narrative) and uses a training set, which is generated automatically.

In this study [8], Naive Bayes and Recurrent Neural Network (RNN) were used to accomplish multilingual sentiment analysis. The translation of multilingual tweets into English was done using the Google Translator API. The outcomes show that RNN outperformed Naive Bayes with 85.34% compared to 77.21%.

The VADER model preserves (and even enhances) the advantages of conventional sentiment lexicons such as LIWC: it is more comprehensive, but equally easy to examine, comprehend, apply rapidly (without requiring a lot of training or learning), and readily expandable. The VADER sentiment lexicon is human-validated and of gold standard quality, similar to LIWC (although not to some other lexicons or machine learning models) [10].

Sentiment lexicons, which are sets of words with emotional scores, are used to assess if a text has a positive, neutral, or negative bias. These lists of scored words, produced by algorithms or experts, feed into feature engineering for machine learning models and power tools such as VADER. Sentiment lexicons provide a straightforward and comprehensible means of interpreting the emotional tone of text data, which opens up possibilities for opinion mining, social media analysis, and other applications, even though domain-specific modifications and context are essential.

A recent work [11] explores the possibility of predicting future stock returns based on tweets related to presidential elections and NASDAQ-100 companies. Another recent work [12] uses convolutional neural network architecture for emotion identification in Twitter messages. Their approach uses unsupervised learning, whereas we use supervised learning and their accuracy of 55.77% is much lower than ours

RoBERTa, BERT's more intelligent progeny, is more accurate and resilient, which makes it ideal for delving into the intricate realm of sentiment analysis in medicine. Beyond simple labels, it deciphers complex contexts, identifies specific aspects of patient viewpoints, and captures subtle emotional nuances, providing a deeper comprehension of the patient experience and improved treatment and care.

The RoBERTa-IAN model proposed in this paper [13] consists of input layer, semantic extraction layer, interactive attention networks layer and emotional output layer. the RoBERTa model to dynamically encode aspect words and context and then perform sentiment classification prediction.

II. METHEDOLOGY

Maṉa is designed to analyze mental health conditions through user interactions with social media. The system collects and analyzes user comments on social media posts and tweets, using a fine-tuned RoBERTa model to classify sentiment (positive vs. negative). Based on the analysis, the system determines whether additional assessment is needed. If negative sentiment predominates, the system triggers the MaṉaNow mode—a dedicated questioning AI that interacts with the user through a structured assessment flow and ultimately generates a final report along with supportive instructions. The system is split into two main interactive modes:

MaṉaChat: A standard instructional chatbot that responds to user inquiries (e.g., "How can I reduce my stress levels?") using the meta-llama/Llama-3.2-3B-Instruct model. MaṉaNow: A questioning chatbot that is activated when negative sentiment or concerning patterns are detected. This mode uses the deepseek-ai/DeepSeek-R1 model to dynamically ask a series of 21 assessment questions, and at the end, generate a final report with insights and supportive guidance.

The user flow is as follows: • Landing Page: Users see a clear landing page with a "Get Started" button. • Project Overview Page: This page provides details about the project and offers two buttons: o "Get Started with MaṉaChat" o "Get Started with MaṉaNow" • Interactive Chat Interfaces: Depending on the chosen mode: o MaṉaChat UI: for standard chat interactions. o MaṉaNow UI: for guided mental health assessment (questionnaire) and final report generation. • Data Collection & Analysis: The system also supports comment collection via APIs, allowing users to upload/download their comments for further analysis at Project Overview UI Model Descriptions: The above proposed system uses the finetuned RoBERTa namely MHRoberta model for sentiment analysis. MHRoBERTa is a Mental Health Roberta which is regressively finetuned on the Mental Health Dataset which is available on Kaggle dataset resources Purpose: To provide direct, instructional responses to user queries related to mental health (e.g., "How can I reduce my stress levels?"). 22 1.Data Collection The data is collected from various platforms, including social media, Reddit, Twitter, and others. Each entry is labeled with a specific mental health status. The dataset contains statements categorized under one of the following seven mental health statuses: Depression, Suicidal, Anxiety Disorder, Stress, Bipolar Disorder, Personality Disorder 2.Custom Dataset Creation, Handling Class Imbalance & text processing The custom sentiment class is created to load into a model and parallel text data is preprocessed by Roberta tokenizers with calculate the class weights (for loss) and sampling weights (for batch balancing) of train and testing set 3.Model configuration and Training process The model is configured through the adapters with 8 class labels and we added customized sentiment adapter In this step Setting the Optimizers and initializing the necessary parameters such as we set number batches=32 for train loader and batches = 128 for validation loader, epochs = 10 Learning rate = 2e-5 to 3e-5 used to achieve the best accuracy 4.Model evaluation and deployment to huggingface After several Iterations of training the customized model we achieved accuracy upto 76.53%.But it is low due to class imbalance in the dataset and loss value is 0.4326 avg train loss. After evaluation the model is saved during training process in drive and the model is deployed in hugging face models as pretrained transformer finetuned on Roberta base. Which could be used as inference and load to the Maṉa application for mental health detection The above proposed system that uses the Meta LLaMa model meta-llama/Llama-3.2-3B Instruct model and deepseek-ai/DeepSeek-R1 model 23 ➔The meta-llama/Llama-3.2-3B-Instruct model used in MaṉaChat (Instructional AI) ➔ Purpose: To provide direct, instructional responses to user queries related to mental health (e.g., "How can I reduce my stress levels?"). Working Principle of Llama model: • Architecture of LLama: o LLaMA-3.2-3B-Instruct is a 3-billion parameter model fine-tuned for instruction-following tasks. o It belongs to the LLaMA (Large Language Model Meta AI) family developed by Meta (Facebook). o The model is optimized to understand and generate human-like responses based on natural language inputs. • Pre Training Process: o Trained on diverse text sources including web content, books, and research papers. o Fine-tuned using supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF) to make responses more useful, safe, and context-aware. • Actual Inference Process: When a user sends a message (e.g., "How can I reduce stress?"), the model: o The model is Prompt Tuned to give the instruction o Processes the input and understands the user's intent. o Retrieves relevant knowledge from its training data. o Generates a step-by-step response with actionable advice. 24 It follows an instruction-tuned approach, meaning it's specifically designed to answer questions concisely and informatively. ➔ The deepseek-ai/DeepSeek-R1 model used in MaṉaNow (Questioning AI for Mental Health Assessment) ➔Purpose: To ask structured questions to assess the user's mental health and generate a final report. Working Principle of DeepSeek-R1: • Architecture of DeepSeek-R1: o DeepSeek-R1 is a powerful Large Language Model (LLM) designed for deep reasoning, interactive assessments, and report generation. o It supports question-answering, summarization, and structured analysis. o Unlike a standard chatbot, this model is optimized for guided step-by-step questioning. • Pre Training Process: o Trained on instruction-based datasets that include psychological assessments, structured questioning, and long-form reasoning. o Fine-tuned for interactive dialogue where it can adaptively adjust the questions based on user responses. o Uses large-scale reinforcement learning (RL) techniques to

enhance question precision and report generation. 25 • Actual Inference Process: When triggered (e.g., after detecting high negative sentiment in user comments), the system: o Asks a series of mental health-related questions (e.g., "How often do you feel anxious?"). o Analyzes the user's responses dynamically—choosing the next question based on previous answers. o Generates a structured final report summarizing the user's mental health status. o Provides tailored recommendations based on the user's responses.

III. MODEL EVALUTION

A. Sentiment Score Determination:

After the Roberta model is fine-tuned to analyze sentiment on drug reviews, it produces three unique sentiment scores for every input, which correspond to the model's evaluation of positive, negative, and neutral attitudes. It is possible to compute a ratio or weighted combination of these individual values to get the final sentiment score for a particular review.

$$Score = (Pos * P_w) + (Neg * G_w) + (Neu * N_w)$$

Where P_w, G_w, N_w are the weights that correspond to the drug reviews' rating.

B. Evaluation the metrices:

Numerous metrics, such as precision, recall, and F1-score, can be used to evaluate the quality of this model. Recall is the percentage of applicable items that are also part of the collection of recommended items, whereas precision is the percentage of advised items that are truly applicable to the user. The F1-score, on the other hand, takes the harmonic means of precision and recall.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ Score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

IV. RESULT

The model's evaluation metrics were computed. With a precision of 92.45%, the model made accurate analysis for 85.5% of the goods. Additionally, the model's recall of 93% indicates that 94.56% of the helpful recommendations are accurate. Furthermore, the model's F1-score of 93.68% indicates that a large percentage of the predictions made by the models in use were accurate. As a result, the RoBERTa model performs admirably.

In the realm of sentiment analysis on drug reviews, using the optimized RoBERTa model has produced encouraging outcomes. Patient experiences with drugs are comprehensively understood by the model, which is trained to identify positive, negative, and neutral thoughts. Context-aware and more nuanced assessment of the sentiments represented in the reviews is made possible by the determined total sentiment score, which is formed from the subtle interaction of these individual scores.

Using this method improves the interpretability of sentiment analysis findings by providing a more detailed understanding of the emotional tone and satisfaction levels in the population. The preceding discoveries hold immense value for medical practitioners and scholars who aim to assess patients' attitudes towards particular drugs, hence promoting better-informed choices regarding pharmaceutical interventions.

The RoBERTa model stands out as the most balanced choice, with the highest accuracy, precision, recall, and F1 score across all models, indicating strong overall performance. VADER, on the other hand, shows the lowest performance in these metrics, suggesting it may not be as effective for nuanced predictions. Logistic Regression performs reliably with balanced precision and recall, although it doesn't quite reach the level of RoBERTa. Naïve Bayes ranks lower, particularly in recall and F1 score, which may mean it could miss certain instances compared to other models. Lastly, SVM is a strong alternative to Logistic Regression, showing good accuracy and F1 score, though it still trails behind RoBERTa in overall effectiveness.

V. RESULT AND DISCUSSION

The Result Analysis of the Mental Health Detection via Social Media project reveals that achieves above 88% accuracy in classifying sentiments correctly. Precision and Recall: Balanced performance across all sentiment classes, minimizing false positives and false negatives. F1-Score: Ensures a high F1-score by considering both precision and recall, making the system effective in real-world applications. Weighted Sentiment Score: The system incorporates user ratings to refine sentiment classification, leading to improved relevance in insights

VI. CONCLUSION

The Emotion Detection via social media project successfully demonstrates the potentialof leveraging social media data to capture and analyze public mood trends. By collecting, processing, and analysing social media posts, this project is able to extract insights about collective emotions, which can be valuable for various applications, including brand

monitoring, public sentiment analysis, and mental health research.The final results are visualized in the reporting module, allowing users to observemood trends over time and identify any significant mood shifts.

Model	Accuracy	Precision	Recall	F1 Score
MHRoBERTa	0.85	0.92	0.93	0.93
VADER	0.75	0.72	0.76	0.74
Logistic Regression	0.82	0.81	0.83	0.82
Naïve Bayes	0.78	0.77	0.79	0.78
SVM	0.85	0.84	0.86	0.85

The outcomes of this project demonstrate that emotion detection via social media is afeasible and effective approach for real-time sentiment analysis. While the initial phase focuses on building a robust pipeline for mood classification, future enhancements can expand the system's capabilities. Real-time mood monitoring, multi-platform integration, multilingual support, and more sophisticated machine learning techniques can further improve the system's performance and utility.

REFERENCES

[1] Depression, 2020. URL: https://www.who.int/news-room/fact sheets/detail/depression. Accessed on: February 22, 2021

[2] Depression self-help guide: Work through a self-help guide for depression that uses cognitive behavioural (CBT), 2020.https://www.nhsinform.scot/illnesses-and conditions/mental-health/mental-health-self-help-guides/depression-self-help-guide.

[3] Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, Sophia Ananiadou, "MentaLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models" WWW'24,May13–17, 2024, Singapore, Singapore https://doi.org/10.1145/3589334.3648137

[4] Ana-Maria Bucur, "Utilizing ChatGPT Generated Data to Retrieve Depression Symptoms from Social Media" arXiv:2307.02313v2 [cs.CL] 6 Jul 2023 https://doi.org/10.48550/arXiv.2307.02313

[5] Oleksandr Romanovskyi, Nina Pidbutska and Anastasiia Knysh "Elomia Chatbot: the Effectiveness of Artificial Intelligence in the Fight for Mental Health" CEUR-WS.org/vol2870/paper89.pdf

[6] Rakesh C Balabantaray, Mudasir Mohammad, and Nibha Sharma "Multi-Class Twitter Emotion Classification: A New Approach"

[7] Bharat Gaind, Varun Syal, Sneha Padgalwar "Emotion Detection and Analysis on Social Media"

[8] M. Balaji, Dr. N. Yuvaraj "Intelligent Chatbot Model to Enhance the Emotion Detection in social media using Bi-directional Recurrent Neural Network"

[9] Batyrkhan Omarov1, Sergazi Narynov and Zhandos Zhumanov "Artificial Intelligence-Enabled Chatbots in Mental Health: A Systematic Review"

[10] Yichen Wang, Aditya Pal, Atlanta, GA, San Jose, CA "Detecting Emotions in Social Media: A Constrained Optimization Approach"

[11] Balcombe ,"AI Chatbots in Digital Mental Health Luke"

[12] Margarita Rodríguez-Ib´anez, Antonio Cas´anez-Ventura, Pedro-Manuel, Cuenca- Jim´ enez "A review on sentiment analysis from social media platforms"

[13] D. Chisholm, K. Sweeny, P. Sheehan et al., "Scaling–up treatment of depression and anxiety: a global return on investment analysis, Lancet Psychiatry", volume 3, 2016, pp. 415–424

[14] K. Kroenke, R. L. Spitzer, J. B. Williams, "The phq-9: validity of a brief depression severity measure, Journal of general internal medicine" 16 (2001) 606–613.

[15] W. W. Eaton, C. Muntaner, C. Smith, A. Tien, M. Ybarra, "Center for epidemiologic studies depression scale: Review and revision, The use of psychological testing for treatment planning and outcomes assessment" (2004).

[16] M. Hamilton, "A rating scale for depression, Journal of neurology, neurosurgery, and psychiatry" 23 (1960) 56.

[17] M. Trotzek, S. Koitka, C. M. Friedrich, Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences, IEEE Transactions on Knowledge and Data Engineering 32 (2018) 588–601

[18] A.-S. Uban, P. Rosso, Deep learning architectures and strategies for early detection of self-harm and depression level prediction, in: CLEF (Working Notes), volume 2696, 2020, pp. 1–12.

[19] Sairam Balani and Munmun De Choudhury. 2015. Detecting and characterizing mental health related self-disclosure in social media. In Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems. 1373– 1378.

[20] Muskan Garg. 2023. Mental health analysis in social media posts: a survey.Archives of Computational Methods in Engineering 30, 3 (2023), 1819–1842.vv

[21] Muskan Garg, Chandni Saxena, Sriparna Saha, Veena Krishnan, Ruchi Joshi, and Vijay Mago. 2022. CAMS: An Annotated Corpus for Causal Analysis of Mental Health Issues in Social Media Posts. In Proceedings of the Thirteenth Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, 6387–6396. https://aclanthology.org/2022.lrec-1.686

[22] ShaoxiongJi, TianlinZhang, LunaAnsari, JieFu,PrayagTiwari and ErikCambria. 2022. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In Proceedings of the Thirteenth Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, 7184 7190. https://aclanthology.org/2022.lrec-1.778

[23] Matthew Louis Mauriello, Thierry Lincoln, Grace Hon, Dorien Simon, Dan Jurafsky, and Pablo Paredes. 2021. Sad: A stress annotated dataset for recognizing everyday stressors in sms-like conversational systems. In Extended abstracts of the 2021 CHI conference on human factors in computing systems. 1–7

[24] MSVPJ SATHVIK and Muskan Garg. 2023. MULTIWD: Multiple Wellness Di mensions in Social Media Posts. (2023).

[25] Jennifer Nicholas, Sandersan Onie, and Mark E Larsen. 2020. Ethics and privacy in social media research for mental health. Current psychiatry reports 22 (2020), 1–7.

[26] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and

Hannaneh Hajishirzi. 2023. Self-Instruct: Aligning Lan guage Models with Self-Generated Instructions. In Proceedings of the 61st Annual

[27] Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In Proceedings of the first ACL workshop on ethics in natural language processing. 94–102.

[27] Amy Bruckman. 2002. Studying the amateur artist: A perspective on disguising data collected in human subjects research on the Internet. Ethics and Information Technology 4 (2002), 217–231.