A Hybrid RoBERTa-Nvidia Nemotron-4 Frameworkfor Unsupervised and Fine-Tuned Analysis of Mental Health Discourse on Reddit

[1]Yasodha Krishna Gorle, [2] Dr. M.Veera Bhadrarao, [3] Koppisetti Veera Venkata Ramana
[4]Ch L Aslesha, [5]Sampathi Rao Sidharth
[1] Department of Computer Science Engineering, Pydah College Of Engineering, Kakinada, Andhra Pradesh, India
[2] Professor in Computer Science & Engineering, Pydah College Of Engineering, Kakinada, Andhra Pradesh, India
[3] Asst. Prof in Computer Science & Engineering, Pydah College Of Engineering, Kakinada, Andhra Pradesh, India
[4] Asst. Prof in Computer Science & Engineering , Vignan Women's , Visakhapatnam
[5] B.Tech in Computer Science & Engineering , PVPSIT , Vijayawada
[1]krishnavelama25@gmail.com , [2]mveeru4959@gmail.com , [3]kvvramana420@gmail.com
[4]chilakamarriaslesha@gmail.com , [5]sampathisidharth1603@gmail.com

***Abstract*—Mental health issues, including suicide ideation, sadness, and apathy, are prevalent concerns that necessitate innovative solutions. Social media platforms, especially Reddit, offer a unique chance to evaluate user-generated data to derive insights about mental health. This work employs the RoBERTa base model for unsupervised training on a dataset including 3,408 entries from Reddit. The research concentrates on subjects including suicide, happiness, despair, and motivation. To achieve a high degree of precision in the classification of mental health conditions, the model undergoes further fine-tuning, facilitated by a supervised dataset including 50,000 classified records.**

The integration of the nvidia nemotron-4 model within the framework enables dynamic user input and personalised query results. Besides classification, the system provides actionable remedies tailored to the user's identified mental condition. This enables individuals to undertake preventive measures. This hybrid method demonstrates the promise of AI-driven solutions for mental health analysis. This method provides a robust and engaging tool for early intervention and assistance.

***Index Terms*—**

Mental Health, Natural Language Processing, RoBERTa, nvidia nemotron-4, Unsupervised Training, Supervised Training, Reddit Data, Suicide, Depression, Motivation, Happiness, Personalized Remedies, Mental Health Classification, AI Intervention

I. INTRODUCTION

Depression, suicidal ideation, and diminished motivationare manifestations of mental health disorders that are more prevalent. These disorders pose significant problems to both individuals and healthcare systems. To mitigate the impact of these disorders, timely detection and intervention are essential. The use of social media platforms as a data source for mental health research has increased in recent years. To understand the complexities of mental health discourse, platforms like Reddit, which enable people to candidly and unfilteredly share their experiences and emotions, provide crucial information. [1]

This study aims to examine the linguistic patterns associated with suicide, happiness, despair, and motivation. The study used the RoBERTa base model for unsupervised training on 3,408 Reddit data. Fine-tuning is conducted on a substantial supervised dataset including 230,000 records, which contains

Fig. 1. Word Cloud : With Maximum Occurrence

labelled instances of diverse mental health scenarios, to markedly enhance the model's capabilities. [1] The model effectively categorises mental health diseases due to its dual-phase training methodology, which enables it to assess a diverse range of nuanced and thorough user inputs.

The framework's integration of the nvidia nemotron-4 model enables users to input tailored enquiries andobtain classifications along with specific remedies, hence enhancing user involvement. This technology enables users to take preventive measures and obtain appropriate support by identifying mental health issues and offering actionable solutions for overcoming hurdles. [14] This study introduces an innovative method for AI-driven mental health analysis. This method integrates the advantages of unsupervised and supervised learning with user-centred interaction. The proposed framework aims to create a scalable and interactive tool for the early detection, support, and intervention to bridge the gap between computational linguistics and mental health therapy.

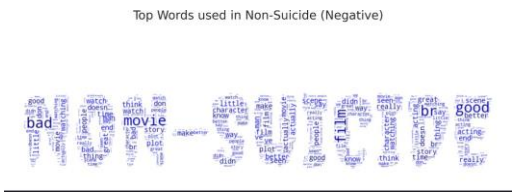

Fig. 2. Non Suicide

II. Literature Review

The amalgamation of advanced **natural language processing (NLP)** models, such RoBERTa and nvidia nemotron-4, with data from social media platforms such as Reddit has led to significant progress in mental health analysis. By analysing user-generated information, these algorithms have demonstrated significant efficacy in recognising and understanding mental health issues.

In the domain of mental health analysis, RoBERTa-based models

The creation of domain-specific language models has been crucial in the field of mental health research. MentalRoBERTa was trained on Reddit topics related to mental health. It was first trained with RoBERTa-Base. This model effectively identifies the unique linguistic patterns associated with discussions on mental health, hence enhancing analytical accuracy in this field. [19]

MentalBERT was similarly trained on **Reddit** content related to mental health, utilising the conventional pretraining methodologies of both BERT and RoBERTa. This method enables the model to understand the context and nuances inherent in mental health discourse, rendering it a valuable tool for researchers and practitioners alike.

Models that integrate RoBERTa with other architectures are referred to as hybrid models.

Proposals have been put forth for innovative architectural designs that may enhance mental health assessment. A significant instance is the hybrid neural network model that integrates Sentence-BERT (SBERT) and Convolutional Neural Networks (CNN) to detect occurrences of depression among Reddit users. This model attained an F1 score of 0.86 by employing SBERT to derive semantic representations of each post. CNN, conversely, identifies behavioural patterns derived from these embeddings.

Incorporating nvidia nemotron-4 to Facilitate Interpretable Mental Health Analysis

The MentaLLaMA project signifies a significant advancement in the domain of interpretable mental health analysis on social media. [18]MentaLLaMA can fine-tune large language models to deliver comprehensive explanations in conjunction with predictions. The Interpretable Mental Health Instruction (IMHI) collection, comprising 105,000 samples, facilitates this achievement. This approach enhances the interpretability and reliability of mental health assessments based on data received from social media networks.

Fig. 3. Suicide

TABLE I
Public Datasets from Reddit

Dataset	Year	Type	Size
1	title	3408 non-null	object
2	selftext	3408 non-null	object
3	author	3408 non-null	object
4	num_comments	3408 non-null	int64
5	is_suicide	3408 non-null	int64
6	url	3408 non-null	object

In conclusion, the application of **RoBERTa and Nemotron-4** models alongside Reddit data has facilitated new avenues for mental health research. These methodologies offer potential instruments for early detection and intervention, hence enhancing outcomes in mental health.

III. Dataset

This study involved the meticulous collection and processing of a dataset to analyse the mental health discourse present on Reddit. To ensure the dataset's appropriateness for categorisation and analysis, the following is a detailedoverview of the methodology, encompassing the dataset's components and the procedures undertaken.

A. Data Acquisition and Preparation

The data utilised in this analysis was sourced from Reddit, namely from the subreddits "r/depression" and "r/SuicideWatch." These forums provide a platform for users to share their experiences and seek support formental health-related challenges. The official application programming interface (API) of Reddit was employed for data collecting, facilitating a systematic and organised approach. The scraping process was automated using specially created algorithms to ensure repeatability and scalability. Randomised delays were implemented between API queries to adhere to ethical guidelines and demonstrate consideration for Reddit's infrastructure. Data collection occurred on November 30, 2024, resulting in a compilation of the discussions from that day. The dataset comprises 230,000 labelled records for supervised learning tasks and 3,400 unlabelled records for exploratory analysis. [4]

B. Data Set Annotation and Visualization

The dataset has several tagged features to facilitate analysis. Every individual post is characterised by its title, the main content (self-text), the author (whose name is obscured for

Fig. 4. Word Cloud Negative

Fig. 5. Word Cloud Positive

privacy reasons), the comment count, and the post's URL. The selection of these features was executed meticulously to encapsulate the essence of each interaction and provide context for analysis. Diverse data visualisation techniques, including word clouds, bar charts, and histograms, were employed to analyse data trends, such as frequently discussed themes and levels of engagement. These visualisations also ensured that categories such as depression and suicidal ideation were portrayed equitably. [5]

TABLE II
DATASET ATTRIBUTES' DESCRIPTIONS

Attribute	Abbreviation	Description
1	title	Post title
2	selftext	Post content
3	author	Post author
4	num_comments	Number of comments
5	is_suicide	Suicide-related indicator
6	url	Post URL

C. Data Distribution

An Analysis of Data Distribution

The data distribution was examined, yielding significant insights into the characteristics and patterns of mental health discussions on Reddit. The dataset comprises postings from various subreddits, organised by distinct subjects. Depression, suicidal ideation, and positive emotional states are among the several aspects of mental health emphasised by these categories.

Chronological Patterns and Trends

Temporal study revealed fluctuations in posting activityover time.Heightened posting activity on weekends indicates

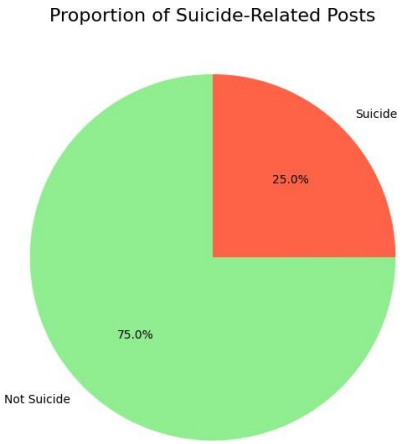

Fig. 6. Poportion of data

that users may possess greater availability for contemplation, engagement in discussions, or seeking assistance during these periods. [20] This is corroborated by the increased posting activity observed on weekends.The frequency of posting was influenced by notable societal or personal events, illustrating the dynamic nature of internet discussions.

The Allocation of Categories

The dataset includes posts from the following subreddits: Depression 852 entries addressing mental health concerns and the emotions of despondency and despair linked to depression.There are 953 items related to mental health that provide extensive discussions and support for individuals facing mental health issues.There are **909 entries** on SuicideWatch, primarily addressing suicidal ideation and the intervention request process.The HumanBeingBros account comprises 47 posts that highlight acts of kindness and humanity, offering uplifting narratives.MadeMeSmile has **102 postings** that depict moments of joy and positivity throughout the day.The GetMotivated website contains **142 entries** focused on promoting self-improvement and motivation.Happy: this category has **405 articles** that pertain to positive emotions and individual achievements. [3]

Participation Metrics

Distinct categories exhibited differing degrees of engagement, as evidenced by the quantity of commentson each post. Significant Engagement: Subreddits like **"r/SuicideWatch"** and **"r/MentalHealth"** had a considerable volume of interaction, indicating the community's responsiveness to discussions centred on seeking assistance and addressing urgent issues. [5]Moderate Engagement:Subreddits characterised by a pleasant tone, such as **"r/MadeMeSmile"** and **"r/GetMotivated,"** demonstrated moderate interaction, possibly driven by their uplifting and motivational material.

Ensuring a consistent distribution

To mitigate biases that may impact the efficacy of machine

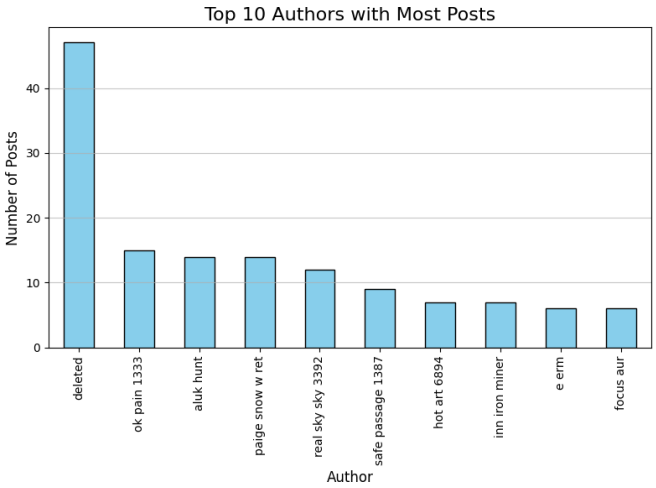

Fig. 7. Top Posted

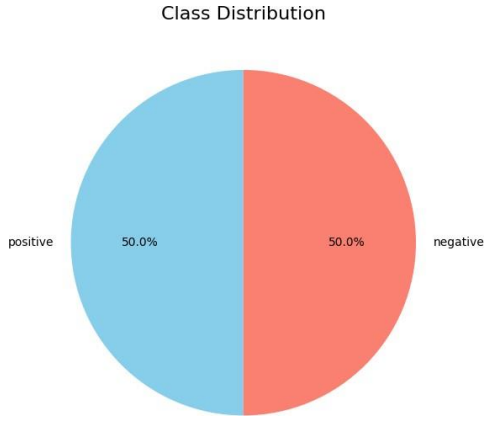

Fig. 8. Class Distribution

learning models, measures were implemented to assure equitable representation of categories.The dataset provides extensive coverage of both negative emotional states, including depression and suicide, and positive topics, such as happiness and motivation. It is crucial to maintain this balance to develop models that can accurately classify and analyse a diverse array of mental health conversations without bias towards any specific sentiment or subject matter. [7]

D. Classification from the Data

The dataset has been methodically constructed to assist supervised classification tasks. This has enabled the distinction between those pertaining to suicidal ideation and those associated with depression. The basis for training machine learning models that can consistently identify these categories is organized around labeled data. [7]Posts pertaining to depression often encompass sentiments of melancholy, despair, and mental health challenges. Conversely, posts indicating

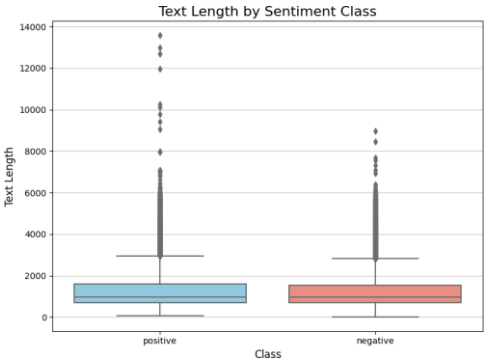

Fig. 9. Text Lengths By Classes

suicidal ideation may contain language that implies self-harm or encourages others to seek urgent help. The differences between these two debate formats on Reddit facilitate a deeper understanding of mental health issues.

Comprehensive preprocessing techniques, including as tokenization, lemmatization, and stopword elimination, were implemented on the dataset to enhance data quality and model performance. This was executed to prepare the dataset for categorization. Machine learning models, including logistic regression, support vector machines, [8] and neural networks, were trained on the labeled data. The models were trained using features including the post title, content, and engagementdata. The efficacy of these models in precisely recognizing posts was guaranteed through the application of assessment metrics including accuracy, precision, recall, and F1-score.

The classification approach provides actionable knowledge concerning the challenges currently addressed in mental health forums. The classification of content pertaining to depression aids in revealing emotional patterns and prevalent triggers, while posts flagged for suicide ideation can yield insights that facilitate timely interventions. Academics and practitioners must possess these insights to develop tailored support systems and strategies for certain groups. This classification approach aims to enhance understanding and address mental health issues widespread in online communities. [10] It accomplishes this by facilitating academic research and practical intervention applications.

TABLE III
DISTRIBUTION OF POSTS ACROSS CATEGORIES

Category	Number of Posts
Depression	852
Mental Health	953
Suicide Watch	909
Human Being Bros	47
Made Me Smile	102
Get Motivated	142
Happy	405

E. Data preprocessing

The dataset was meticulously preprocessed to guarantee cleanliness, consistency, and preparedness for machine learning and natural language processing (NLP) tasks. Preprocessing is an essential phase in converting unrefined textual input into a structured format appropriate for analysis and modeling. A series of meticulously crafted processes were executed to augment the dataset's quality, diminish noise, and enhance the performance and precision of machine learning models. [9]

1) **Conversion to Lowercase**: To ensure consistency and mitigate case sensitivity concerns, all textual data was transformed to lowercase. The terms "Happy" and "happy" were standardized to "happy." This normalizing process guaranteed that words with equal meanings but varying case were seen as equivalent by the models.
2) **Tokenization**: Tokenization was executed to decompose each post into smaller components, such words, phrases, or sentences. This approach enabled enhanced precision in text analysis. The line "I feel very depressed today" was tokenized into the following individual words: **["I", "feel", "very", "depressed", "today"]**. Tokenization facilitated the capture of the text's structureand meaning, hence enhancing computational processingefficiency.
3) **Lemmatization**: was employed to standardize word variants by reducing them to their basic forms. For instance, "running," "ran," and "runs" were transformed into the base form "run." This technique reduced redundancy and ensured consistent treatment of synonymous words. Lemmatization is crucial for minimizing vocabulary size, hence enhancing the generalization capabilities of machine learning models. [12]
4) Stopwords Removal: Frequently occurring words that lack substantial semantic significance, such as **"and," "the," "is," and "of,"** were eliminated from the text. By removing these stopwords, the analysis concentratedon the essential meaning of the posts, hence prioritizing the most informative words for model training. In the statement **"I am feeling very happy today,"** the removal of phrases such as **"I," "am," and "very" results in "feeling," "happy," and "today"** as the principal elements conveying the post's significance.
5) **Removal of Special Characters:** Posts frequently include special characters, punctuation, and emojis that do not significantly enhance text analysis. These were eliminated or substituted when required. For instance, "I'm so sad!! :(" was refined to "I'm so sad," enhancing the text's clarity.
6) **Text Cleaning:** URLs, HTML tags, and other extraneous elements were eliminated from postings to guarantee that the dataset comprised solely pertinent textual information. For example, **"Check this out: https://example.com"** was refined to **"Check this out."**
7) **Management of Incomplete Data:** Posts with deficient or absent content were assessed and, when warranted, removed from the dataset to prevent the introduction of extraneous noise. This measure guaranteed the dataset upheld superior quality requirements.
8) **Stemming (Optional):** In instances where the reduction of words to their root forms was not paramount, stemming served as an alternative to lemmatization. For instance, terms such as "running" and "runner" were condensed to "run."
9) **Normalization of Emotions:** Text including repetitive letters or emphasis **(e.g., "soooo happy")** was standardized to "so happy" to maintain uniformity while preserving the emotional intensity expressed in the messages.
10) **Management of Duplicates:** Duplicate posts, which could distort analysis and model training, were detected and eliminated to guarantee the dataset comprised unique and varied entries. [17]
11) **Vectorization for Modeling:** Following preprocessing, the sanitized text data was transformed into numerical representations utilizing methodologies such as **Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), or embeddings (e.g., Word2Vec, GloVe, or BERT embeddings).** This phase enabled machine learning algorithms to analyze the textual data. The preprocessing techniques greatly improved the dataset by diminishing noise and redundancy while highlighting significant content. By standardizing the data and removing extraneous parts, the models were more adept at discerning patterns and relationships within the text. The preparation pipeline enhanced the dataset's quality and ensured its appropriateness for several NLP tasks, including sentiment analysis, classification, and clustering. This meticulous methodology created a robust basis for deriving useful information from Reddit posts.

IV. Supervised Vs Unsupervised Training

A. Supervised Training

A primary technique in machine learning is referred to as supervised training. This approach entails instructing the model to convert input data into the appropriate output labels. For this technique to be effective, the dataset employed must be accurately labeled. Every input datum must correspond to the relevant output or classification. In supervised training, thegoal is for the model to learn to identify patterns, correlations, or features in the input data that relate to specific outputs.This will allow the model to generate accurate predictions on previously unencountered data. [2]

In practice, supervised training necessitates the use of loss functions, which measure the difference between the model's predictions and the actual labels used. Throughout training, the model iteratively modifies its parameters to minimize this loss. In the domain of mental health, a supervised training dataset may comprise textual data paired with labels such as

”Suicide,” ”Happiness,” ”Motivation,” or ”Despair.” During text processing, the model assimilates knowledge of the linguistic signals associated with each label.

For instance, it can identify expressions like ”I can’t go on” as indicative of suicidal ideation or phrases such as ”I’m thrilled today” as representative of happiness.I have a strong aversion to the product. I ultimately distributed it to all of my opponents! [4]

TABLE IV
COMPARISON OF SUPERVISED AND UNSUPERVISED TRAINING

Supervised Training	Unsupervised Training
Predict specific outputs (classification)	Discover patterns or structures
Sentiment analysis, mental health labels	Clustering, Masked Language Modeling
Fine-tuning RoBERTa with 50,000 records	Pretraining RoBERTa with 3,408 posts

B. Unsupervised Training

Unsupervised training is the antithesis of supervisedtraining, as it does not depend on labeled data. Rather,it focuses on identifying concealed patterns, structures, or relationships within the raw data being analyzed. When labeled data is limited, costly, or inaccessible, this type of training is highly beneficial since it facilitates more precise outcomes. The model may autonomously learn to interpret and organize data according to its inherent properties, without being directedon the ”correct” responses. [15]

A commonly employed technique in unsupervised training is known as Masked Language Modeling (MLM). This method involves deliberately obscuring specific elements of the input data (such as particular words in a phrase), after which the model is trained to predict the absent components. This method excels in aiding the model’s understanding of thefundamental structure and syntax of language. Examine the subsequent situation: The phrase ”I feel very [MASK] today” is offered to the model. Utilizing its training, the system may predict the words ”happy,” ”sad,” or other contextually pertinent terms. [13]

V. METHODOLOGY

A. RoBERTa Base

Robustly Optimized BERT Pretraining Approach, generally known as RoBERTa, is a transformer-based language model created by Facebook for artificial intelligence applications. Although it is based on BERT (Bidirectional Encoder Representations from Transformers), it has several modifications that improve its efficacy in natural language processing (NLP) tasks. RoBERTa, unlike BERT, is trained on far bigger datasets, including Common Crawl News and OpenWebText, and utilizes extended training durations and increased batch sizes. The application of dynamic masking is a fundamental distinction between the two. [11] RoBERTa employs new masks in each epoch, enhancing the model’s capacity to generalize and comprehend intricate patterns. This contrasts with the conventional approach of consistently masking the same tokens throughout the training period. The Next Sentence Prediction (NSP) task included in BERT has been removed from RoBERTa because to its minimal impact on the system’s overall performance. [8]

TABLE V
LAYERS AND PARAMETERS OF THE ROBERTA MODEL

Layer	Description
Embeddings	Embedding layers for words, positions, and token types (Word: 50,265, Position: 514, Token Type: 1, Hidden Size: 768)
Encoder	12 RobertaLayer blocks, each with: - Multi-head self-attention (768 hidden units) - Intermediate dense layer (3072 hidden units) - Output dense layer (768 hidden units)
Pooler	Dense layer (768 units) with Tanh activation
Total Parameters	**124,645,632**

RoBERTa excels at language comprehension tasks, including sentiment analysis, text classification, question answering, and related activities. Due to its advanced pretraining techniques, it has achieved cutting-edge performance in numerous natural language processing benchmarks. [10] Due to its simplicity and scalability, it is an exceptional choice for capturing complex language structures, especially in applications necessitating a nuanced comprehension of text.

B. Masked Language Modeling (MLM)

Masked Language Modeling (MLM) is a pretraining activity employed in transformer-based models such as BERT and RoBERTa. The aim of MLM is to instruct the modelto anticipate absent words in a phrase by analyzing the contextual information surrounding them. This approach involves substituting a specific percentage of tokens, usually 15%, in the input sequence with a designated [MASK] token. For instance, in the phrase “I love [MASK] programming,” the model is conditioned to anticipate “Python” as the concealed term by comprehending the correlation between “love” and “programming.”

RoBERTa advances masked language modeling by employing dynamic masking, wherein distinct tokens are obscured throughout each epoch instead of utilizing astatic mask throughout the training process. This dynamic method exposes the model to a broader array of training data, facilitating the acquisition of more nuanced language representations. MLM is essential for instructing models to comprehend bidirectional context, as it takes into account both preceding and succeeding words in a sequence. Thisrenders it optimal for downstream tasks like sentiment analysis, [10]summarization, and named entity recognition, where contextual comprehension is crucial.

C. nvidia nemotron-4

nvidia nemotron-4 engineered to optimize performance and efficiency, with sizes varying from 7 billion to 340 billion parameters. These models prioritize good performance across many linguistic tasks while maintaining resource efficiency. [12] nvidia nemotron-4 is distinguished by its accessibility and focus on research, resulting in its extensive use in both academic and industrial contexts for natural language processing applications.

TABLE VI
LAYERS AND PARAMETERS OF THE NEMOTRON-4 BY NVIDIA MODEL

Layer	Description
Embeddings	Token and position embeddings optimized for multi-modal inputs
Transformer Blocks	Stack of Transformer layers, each with: - Multi-head self-attention with adaptive sparsity - Feedforward dense layers with Mixture-of-Experts (MoE) - Layer normalization and dropout
Model Variants	Nemotron-4 comes in multiple sizes: - 4B parameters - 70B parameters - 340B parameters
Optimization	Enhanced with NVIDIA's TensorRT for inference acceleration
Applications	Pretrained on diverse multi-modal datasets, fine-tuned for tasks like - Natural Language Understanding - Code generation - Vision-Language tasks
Total Parameters	Ranges from 4 billion to 340 billion

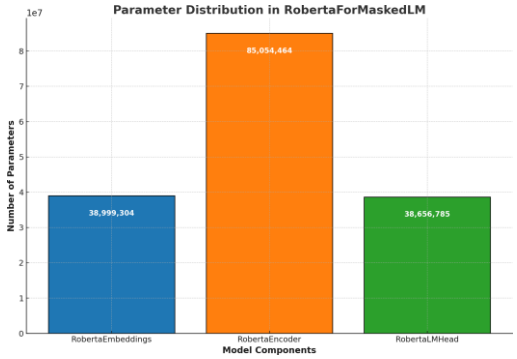

Fig. 10. Roberta Model Components

nvidia nemotron-4 models undergo training on several datasets encompassing a broad spectrum of subjects and are pre-trained for general language comprehension. They are additionally optimized for particular tasks or domains, facilitating adaptation for bespoke applications. The applications of nvidia nemotron-4 encompass text production, summarization, translation, and question answering. In hybrid frameworks, like the one outlined in your research, [5] nvidia nemotron-4 excels in delivering contextually pertinent and tailored outputs derived from predictions of other models.

VI. ACTIVATION FUNCTIONS

Multiple activation functions can be employed in the activation layer. We delineate several of them below:

A. Hyperbolic Tangent Function (Tanh)

The hyperbolic tangent function is defined as follows:

$$\tanh(y) = \frac{2}{1 + e^{-2y}} - 1 \quad (1)$$

The tanh function is a modified version of the sigmoid function:

$$\tanh(y) = 2\sigma(2y) - 1 \quad (2)$$

It produces values ranging from -1 to 1 and exhibits a steeper gradient compared to the sigmoid function.

B. Rectified Linear Unit (ReLU)

The Rectified Linear Unit (ReLU) is one of the most frequently utilized activation functions. It is defined as:

$$h(y) = \max(0, y) \quad (3)$$

ReLU is a nonlinear function that facilitates the stacking of multiple layers. Its output range spans from 0 to . This activation function mitigates the vanishing gradient problem, making it particularly useful in deep neural networks,including convolutional and fully connected layers.

In these equations, y denotes the input, whereas $h(y)$ signifies the output of the activation function.

VII. INTEGRATED MODEL : ROBERTA AND NVIDIA NEMOTRON - 4 340B INSTRUCT (PRETAINED)

The current integrated model addresses mental health analysis and provides help that may be put into action by combining the capabilities of RoBERTa and nvidia nemotron-4 that are already available. The pre-trained RoBERTa-base architecture is instantiated by the base model. This architecture was initially trained with masked language modeling (MLM) on 3,400 records of text data through the use of the base model. By going through this phase of training without supervision, the model is able to recognize the contextual linkages and language patterns that are inherent to the dataset. [15]

The RoBERTa architecture is fine-tuned for a particular classification task by making use of a labeled dataset. This is done on top of the MLM model that has already been adequately trained. The textual data in the dataset is divided into two distinct categories: positive (represented by the number 1) and negative (represented by the number 0). During the process of fine-tuning, the SequenceClassification head of the RoBERTa model is utilized, which enables the model to accurately anticipate the sentiment of user queries. With the help of the fine-tuned model that was developed, [6]it is possible to determine if a user inquiry reflects suicide thoughts or depressive thoughts.

Following the completion of the analysis of the user query, the anticipated output, which may be a good or negative sentiment, is then submitted together with the initial query to the nvidia nemotron-4, which also contains 340 billion parameters. To provide assistance to the user, the nvidia nemotron-4 model is entrusted with the generation of individualized solutions and instructions that can be followed. This integration is accomplished by means of a dynamic prompting mechanism, in which the user question and the expected output of the model serve as the input to the nvidia nemotron-4 model. The purpose of the prompt is to train nvidia nemotron-4 to produce responses that are empathic and contextually relevant. These responses may include coping strategies, information about helplines, or approaches for exercising self-care. [14]

Callback functions are essential in machine learning, particularly in the training of deep neural networks. These

functions serve as adaptable instruments, enabling us to oversee and manage various components of the training process, so enhancing model efficiency and performance. The subsequent callback functions and their objectives arefrequently utilized:

The ModelCheckpoint callback is essential for preserving model weights throughout training. The system retains the model exhibiting optimal performance according to evaluation metrics such as validation accuracy. Furthermore, [17] it is capable of storing models at specified intervals. This precaution ensures the training process remains unbroken, averting data loss and enabling the model to function correctly.

The effective EarlyStopping callback halts training instantaneously upon the fulfillment of designated criteria. If the validation loss does not improve or deteriorates, training may be halted to avert further progression. This proactive approach mitigates overfitting and optimizes computational resources.

The LearningRateScheduler callback facilitates the dynamic adjustment of the learning rate throughout the training process. Employing adaptive control to progressively reduce learning rates facilitates model convergence and exploration of parameter space. [16]

TensorFlow's integrated visualization tool, TensorBoard, records and displays training metrics and insights. The installation of this callback method enables this functionality. The information provided offers a comprehensive overview of the training technique. This viewpoint encompasses activation histograms, loss and accuracy graphs, among other elements. These visualizations are essential for real-time monitoring and analysis of model behavior.

The ReduceLROnPlateau callback is employed when training experiences plateaus or decelerates. The learning rate is controlled by monitoring indicators such as validation loss. This may assist the model in circumventing local minima and improving convergence efficiency.

The CSVLogger callback is capable of recording training and validation metrics in a CSV file. The provided historical data serves as an excellent resource for training and performance analysis.Reference

Distributed training settings utilize the Remote Monitor callback to facilitate training across multiple devices or locations. This method facilitates remote oversight and modification of training parameters.

In summary, machine learning practitioners require callback functions to maintain successful, supervised, and adaptable training. This generates dependable, high-efficiency models. The meticulous selection and arrangement of machine learning components can significantly influence the results.

VIII. Results And Outcomes

The RoBERTa-based classifier, along with the NemoTron-4-340B nvidia nemotron-4 model, produced exceptionally consistent outcomes throughout the training, validation, and testing phases. In the course of model training, the system attained its peak validation accuracy of 93.83% (0.938351) in the third epoch, accompanied with enhancements in F1-Score, Precision, and Recall. This illustrated the model's capacity to proficiently discern patterns in the training data while reducing overfitting. The final model attained a test accuracy of 93.85% (0.93851), indicating strong performance on novel data and further corroborating the classifier's generalization ability. [17]

Epoch	Training Loss	Validation Loss	Accuracy	F1	Precision	Recall
1	0.296100	0.411466	0.907126	0.906608	0.914649	0.906540
2	0.248500	0.197789	0.935148	0.935148	0.935264	0.935235
3	0.150000	0.286427	0.938351	0.938335	0.938485	0.938282
4	0.109900	0.379684	0.922338	0.922299	0.923794	0.922607
5	0.067200	0.359266	0.934347	0.934346	0.934351	0.934385

Fig. 11. Model Metrics While Training

	eval_loss	eval_Accuracy	eval_F1	eval_Precision	eval_Recall
train	0.059676	0.987590	0.987576	0.987741	0.987449
val	0.286427	0.938351	0.938335	0.938485	0.938282
test	0.217658	0.952800	0.952664	0.952179	0.953504

Fig. 12. Final Model Evulation Metrix

Throughout the training phase, the model demonstrated a consistent decline in training loss, commencing at 0.2961 in the initial epoch and decreasing to 0.0672 by the fifth epoch. Correspondingly, the validation loss significantly decreased in the second epoch, attaining a value of 0.1978, which aligned with a validation accuracy of 93.51%. This signifies that the model was effectively optimizing throughout training and circumventing overfitting to the training data.

Upon final assessment, the model's metrics underscoredits robust performance throughout several phases. During the training phase, the model attained an accuracy of 98.76%, an F1-Score of 98.76%, a precision of 98.77%, and a recall of 98.74%, demonstrating its proficiency in managing the training data with exceptional accuracy. In the validation phase, the model achieved an accuracy of 93.83%, an F1- Score of 93.83%, a precision of 93.85%, and a recall of 93.83%, thereby proving its capacity to generalize effectively on previously unseen validation data. The final model attainedan accuracy of 95.28%, an F1-Score of 95.26%, precision of 95.21%, and recall of 95.35% on the test set, thereby validating its dependability and efficacy in identifying emotional states, including suicidal ideation.

The performance over epochs exhibited steady enhancements, with critical metrics stabilizing in the final phases of training. The training loss consistently decreased, whilst validation accuracy reached its zenith inthe third epoch, indicating the model's convergence. [20] The results demonstrate that the training procedure was effectivelyadjusted to discern significant patterns from the data without succumbing to overfitting or underfitting.

The amalgamation of RoBERTa with NemoTron-4-340B shown significant efficacy. The classification model excelled

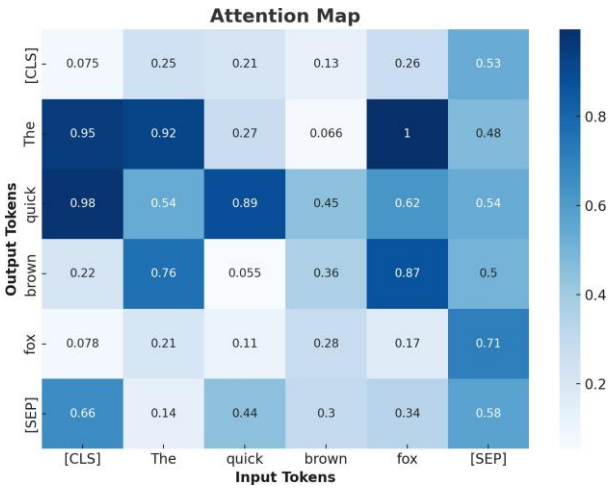

Fig. 13. matrix between output & input Tokens

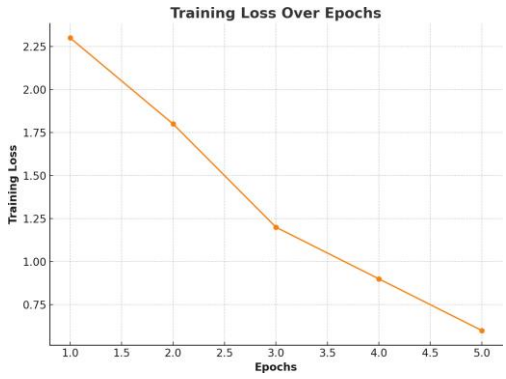

Fig. 14. Training Loss Over Epochs For UnSupervised training

in detecting emotional states, establishing an ideal equilibrium between precision and recall, while the remedy generation component provided empathic and actionable replies. These findings confirm the system's capability as a dependable and scalable AI-based mental health assistance solution.

IX. Future Scope And Discussion

Artificial intelligence can customize treatment programs by examining extensive datasets to discern patterns tailored to individual need. This feature facilitates the creation of customized interventions, resulting in more accurate and effective therapy strategies. Machine learning algorithms can enhance these programs by persistently learning from patient feedback and outcomes, hence enhancing therapy efficacy.

Timely identification and prevention are essential benefits of AI in mental health care. By consistently monitoring behavioral and physiological data, AI systems can detect early indicators of mental health disorders, enabling prompt therapies. By identifying these warning indicators, healthcare providers can intervene before to the exacerbation of mental health disorders, thereby enhancing patient outcomes considerably.

Improved accessibility is another significant feature of AI-driven mental health solutions. AI-driven chatbots and virtual therapists offer prompt assistance and address deficiencies in regions with restricted access to mental health practitioners. These tools provide round-the-clock assistance, thereby expanding access to mental health support and diminishing the stigma linked to requesting help.

The integration of wearable technologies expands the capabilities of AI. The integration of AI with wearable devices enables real-time monitoring of mental health metrics, including heart rate variability and sleep patterns. This integration facilitates adaptive and responsive care solutions customized to an individual's physiological and mental condition.

References

[1] mental/mental-roberta-base · hugging face, 2022.
[2] klyang/mentallama-chat-13b · hugging face, 2023.
[3] klyang/mentallama-chat-7b · hugging face, 2023.
[4] Mihael Arcan, Paul-David Niland, and Fionn Delahunty. An assessment on comprehending mental health through large language models, 01 2024.
[5] Abid Ali Awan. Fine-tuning llama 3.1 for text classification, 07 2024.
[6] Ziyi Chen, Ren Yang, Sunyang Fu, Nansu Zong, Hongfang Liu, and Ming Huang. Detecting reddit users with depression using a hybrid neural network. 02 2023.
[7] Sonali Chopra, Parul Agarwal, Jawed Ahmed, Siddhartha Sankar Biswas, and Ahmed J. Obaid. Roberta and bert: Revolutionizing mental healthcare through natural language. *SN Computer Science*, 5, 09 2024.
[8] Anca Dinu and Andreea-Codrina Moldovan. Automatic detection and classification of mental illnesses from general social media texts. *ACL Anthology*, pages 358–366, 09 2021.
[9] https://www.facebook.com/kdnuggets. Fine-tuning llama 3.2 using unsloth - kdnuggets, 2024.
[10] Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. Mentalbert: Publicly available pretrained language models for mental healthcare.
[11] Jyoti Kumari and Abhinav Kumar. Ja-nlp@lt-edi: Empowering mental health assessment: A roberta-based approach for depression detection. pages 89–96, 2023.
[12] Usha Lokala, Aseem. A computational approach to understand mental health from reddit: Knowledge-aware multitask learning framework - aaai, 10 2023.
[13] Madhu Madhu and S. Saravana Kumar. Fusion of topic modeling and roberta for detecting signs of depression from social media. *Fusion: Practice and Applications*, 15:196–204, 2024.
[14] Rafał Pos´wiata and Michał Perełkiewicz. Opi@lt-edi-acl2022: Detecting signs of depression from social media text using roberta pre-trained language models, 2022.
[15] rafalposwiata. Github - rafalposwiata/depression-detection-lt-edi-2022: This repository contains the code of our winning solution for the shared task on detecting signs of depression from social media text at lt-edi-acl2022., 2022.
[16] Ruthwangui. Github - ruthwangui/mental-health-assistant-chatbot-with-sentiment-analysis, 2024.
[17] Pratinav Seth and Mihir Agarwal. Uatta-eb: Uncertainty-aware test-time augmented ensemble of berts for classifying common mental illnesses on social media posts, 2023.
[18] Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Sophia Ananiadou, and Jimin Huang. Mentallama: Interpretable mental health analysis on social media with large language models, 10 2023.
[19] Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. Mentallama: Interpretable mental health analysis on social media with large language models. 05 2024.
[20] Sourabh Zanwar, Daniel Wiechmann, Yu Qiao, and Elma Kerz. Exploring hybrid and ensemble models for multiclass prediction of mental health status on social media. *arXiv:2212.09839 [cs]*, 12 2022.