

A Next-Generation NLP Framework for Psychological Behavior Analysis Based on State-of-the-art Language Model

Mohit Kumar

Amity University

Ashwani Kumar

Bennett University

Avinash Kumar Sharma

Sharda University

Nishant Gupta

ITS engineering College

Achyut Shankar

Bennett University

Gautam Kumar

`gautam.kumar@jaipur.manipal.edu`

Manipal University Jaipur

Article

Keywords: Psychological Behavior, Large Language Model, Interpretable, Transformer, Encoder, Mental Health

Posted Date: April 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9158982/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Psychological behavior analysis is critical in mental health diagnosis, but typical machine learning approaches frequently struggle to appropriately read complex psychological cues. This research fills that gap by presenting a new framework that combines Large Language Models with psychological theory mapping to improve behavior classification. The method tries to close the gap between machine learning performance and clinical relevance by adding well-known psychological theories, resulting in high accuracy and interpretable findings for real-world applications. The proposed model has been assessed on different types of standard and benchmarked dataset where the accomplished outcome exhibited notable performance gain of proposed model in contrast to existing system. The proposed model is recorded with 93.1% accuracy and 92.9% of F1-Score in contrast to several potential approaches e.g. Support Vector Machine, Random Forest, Convolution Neural Network. The outcome states the reliability of the model towards identifying the psychological behaviour offering interpretable and faster clinical prediction leveraged for real-world capabilities. The study contributes to a unified framework that simplifies the diagnostic process of mental health with elevated quantified performance being recorded.

1. Introduction

There is an increasing attention towards usage of Natural Language Processing (NLP) because of its capability for studying psychological actions especially in the area of mental health [1]. Text-based system are highly used for diagnosis issues pertaining to mental health as a part of psychological behaviour analysis. This can eventually lead to an appropriate identification of behavioral traits, cognitive process, cognitive states [2]-[4]. On the contrary, the existing mechanism are found to be under capable towards extracting the essential traits of human behaviour especially when encountered with personality disorder, anxiety, depression, etc. Apart from this there is an absence of interpretability or any direct link to the identified theory of psychology that are essential for decision making. Various emerging areas towards mental health diagnosis has been opened up by adopting Large Language Models (LLM) viz. Generative Pre-trained Transformer-3 (GPT-3), Bidirectional Encoder Representations from Transformers (BERT), etc [5]-[8]. There is a report of successful attempt upon adoption of NLP based task in this regards which is mainly due to their ability to capture subtle changes as well as context in human language related to task viz. text generation, question answering, sentiment analysis. Irrespective of reports of various successful implementation of NLP, such models are found to be recently started to be implemented in mental health based studies. Such studies demands determining and interpreting emotional states and behavioral patterns [9][10].

An increased usage of Artificial Intelligence (AI) towards studying mental health has been witnessed in various literatures claiming of enhancing diagnostic performance [11]. For a given clinical configurations, an appropriate with faster mental health assessment can enhance the results by permitting for highly particular diagnosis and treatment regimen. Unfortunately, current approaches have higher dependencies of manual investigation or they do not possess any anticipated theoretical depth that is necessary for interpreting the outcome. The current state of work consists of various impediments e.g.,

usage of shallow machine learning (ML) techniques or rule-based methods. Such techniques eventually fails to extract the subtle cognitive distortions and emotional states observed within the text. This yields to sub-optimal performance towards identifying the psychological behavior especially when used with complex form of emotional data.

Along with accomplishing higher classification accuracy, it is also necessary to understand the standard theory of mental health where behaviour plays a vital role. There are various mechanism that do not connect to standard behavior to standard psychological theory (DSM-5 criteria or Beck's cognitive triad) that restricts their adoption in clinical usage. Majority of the deep learning models work on principle of black boxes that doesn't offer any form of interpretation nor facilitate any explainability. This poses a challenges in front of physician to actually trust the outcomes of AI models. It is very much necessary to carry out real-time analysis especially when it relates to digital mental health demanding speedy responsiveness. It has been noted that existing models of AI are also found unable to handle this issue.

Therefore, the proposed study offers a solution by introducing a distinct computational and analytical framework that is meant for evaluating the psychological activities using language modelling. The proposed model targets towards precisely representing psychological activities followed by connecting them to the conventional standard psychological theories. The notion is to facilitate a comprehensive visualization of the mental state of an individual. The agenda of the system is to contribute actionable findings, interpretable, and clinically meaningful in practical world for bridging the trade-off between mental health study and NLP usage that prior work has not addressed. The proposed model adopts deep contextual understanding in order to detection emotional traits and classify behaviour while coherent explanation is generated by the language model. The contribution of the study are as follow:

1. The study unifies bidirectional encoder and language model for categorizing behaviour and theoretical mapping for efficient analysis of behaviour.
2. The model categorizes actions along with mapping them to identified standard psychological framework towards comprehensive clinical practices contextually.
3. The proposed study integrates psychological theory with behavioral classification with contextual insights towards assisting in treatment planning, therapy, and diagnosis.

The organization of this manuscript is as follows: Section 2 has discussed about existing prevalent techniques while elaboration of adopted solution is presented in Section 3 in form of research methodology. Elaboration of an accomplished study findings are carried out in Section 4 while Section 5 presents conclusion of this paper.

2. Review of Literature

This section presents discussion of various studies being carried out to understand the methodologies being adopted towards investigating various form of psychological behaviour. It is noted that ML and NLP has a significant role in doing so while its usage is being proliferated among existing literatures.

There were various types of approaches that have been studied in past decade while these studies will be investigated in this section as follows:

2.1. Traditional Approaches to Psychological Behavior Analysis

It has been noted that analysis of historical psychological behaviour has been usually carried out adopting paper-based questionnaires, structured interviews, clinical observation [12]. It can be noted that all these methods are manually performed. Unfortunately, all these mechanisms are highly prone to human artifacts or it could be highly subjective as well as time consuming. This will eventually result in inconsistent diagnosis. There are wider adoption of two essential psychometric test for assessing mental health viz. Beck Depression Inventory (BDI) [13] and Minnesota Multiphasic Personality Inventory (MMPI) [14]. Despite of these assessment tool to be highly standardized yet they too have shortcomings which is related to their being static nature and has highly restricted flexibility for identifying subtle or dynamic psychological activities especially in real-world environment.

Currently, there are various researchers who have used ML-based framework for automating the identification process of psychological behaviour. Some of the research work has noted to be heavily dependent on adopting supervised ML approaches like Random Forest or Support Vector Machine (SVM) for determining psychological states depending upon the organized input of data either inform of speech or text [15]. Such form of computational model demands larger dataset along with labelling to be used for training. Such availability of labelled dataset for certain specific mental disorder is questionable as well as highly challenging. Apart from this, such frameworks also have a shortcomings related to absence of interpretability and overfitting. These feature deem such methods unsuitable towards clinical usage.

2.2. Emergence of NLP and Learning Models

Apart from typical ML models, the proliferated usage of Deep Learning (DL) models as well as architecture based on neural network has been reported. For an example, there is an increasing usage of Convolution Neural Network (CNN) as well Recurrent Neural Network (RNN) proving more adoption of data-driven methodologies towards investigating mental health. Various psychological applications have been designed based on this concept where identification of emotions and sentiment analysis has been carried out in area of NLP. The work carried out by Udding et al. [16] has investigated identification of depression followed by classifying text obtained from social media post using RNN. Adoption of CNN has been reported in work of Kerasiotis et al. [17] where the states of emotions have been identified in similar way.

Existing studies have also witnessed rise of transformer based frameworks e.g. GPT and BERT as an advanced form of NLP meant for revolutionizing the capability of extracting textual dependencies of long term. Similarly, approaches like BERT uses bidirectional approaches for understanding various context that is mainly associated with words and sentences present in textual corpus. This adoption potentially

benefits operations related to entity recognition, question answering, sentiment analysis [18]-[20]. The GPT-3 language model facilitated by the OpenAI has leveraged the NLP to a higher extent where human-like writing are yielded using generative technique. This facilitates GPT-3 to understand various ranges of text while the system also generates a contextually relevant responses [21]-[23].

Adoption of such advanced analytical processing language model is noted to be increasingly adopting in analysis of psychological behaviour. The utilization of BERT model has been reported in work of Reitberg et al. [24] which is carried out towards categorization of clinical text. There are also studies where GPT-3 has been used for facilitating customized response depending upon the emotions of the user [25]-[27]. Various feasibility of sophisticated frameworks further evolves towards identifying the subtle changes in data for enhancing analysis system towards psychological behaviour.

2.3. Behavior Classification and Psychological Theory Mapping

Although, there are various advantageous features and higher scope for adopting NLP techniques towards classification of behaviour, they are also associated with absence of potential theoretical foundation that is much needed in clinical usage. Various psychological actions are demanded to be contextualized in order to obtain useful information by the clinicians. However, most NLP-based systems do not contain psychological theories, therefore the outcomes are insufficiently interpretable or clinically relevant.

Psychological theories utilized in psychological behavior analysis include CBT [28] and the Big Five personality traits [29]. Beck's Cognitive Triad [30], which emphasizes the interplay of negative ideas, feelings, and behaviors, provides an important framework for understanding illnesses such as depression and anxiety. The Big Five personality traits—openness, conscientiousness, extraversion, agreeableness, and neuroticism—are widely employed in personality psychology and provide a solid foundation for explaining individual differences in behavior. A few studies have attempted to close the gap by connecting psychological behavior classifications to theoretical models. Hutto et al. [31] provided a methodology that assessed text data using NLP models and related the identified behaviors to DSM-5 diagnostic criteria. However, these approaches are still in their early phases, and much more effort is required to completely integrate theoretical mapping with real-time behavioral research.

2.4. Recent theory-aware and Psychologically Grounded Transformers

There are also various studies where distinct psychological theories have been integrated with transformer-based models with the sole intention of enhancing the domain relevance and interpretability. The evolving model, e.g. MentalHealth-RoBERTa [32], incorporates lexical biases with a specific domain, and it is trained on Reddit forums associated with health. Another evolving model known as PsyQA-BERT [33] is capable of aligning the objectives of queries and responses solely guided by the framework of CBT with a therapeutic intent. This framework contributes towards establishing a correlation of

psychological entities with data representation to prove its significant enrichment for the transformer-based model. It also facilitates outcomes with contextual awareness capable of aligning itself with diagnosis methods in mental health. However, different from these models, the proposed study incorporates psychological theory mapping with post-classification that offers extensive modularity and adaptability to the novel LLM without any involvement of training, and yet it facilitates interpretability anchored to behavioral studies.

2.5. Multimodal mental health models

Apart from all the text-based frameworks, there are also various evolving research works towards a multimodal framework towards assessing mental health using various types of data, viz., physiological data, video, speech, etc. The Cross-Modal Contrastive Learning (CMCL) [34] and Multimodal Bi-Transformers (MMBT) [35] have been reported of enhanced outcomes towards the detection of emotion and behavior by fusing visual or voice signals or text data. The fusion framework of DAIC-WOZ [36] is also known for integrating acoustic attributes like prosody for improving the accuracy performance which detecting depression. Interestingly, all these models are still at a nascent level and demand more rigorous assessment to yield a result that is applicable in the practical field. This is exactly the context of applicability of the practical field that is the emphasis of the proposed study. Although our current model emphasizes textual information, its outcome is more practically applicable in contrast to all evolving approaches.

2.6. Benchmarking landscape and state-of-the-Art comparison

The work of WASSA-2023 and CLPsych 2021 has been subjected to a benchmarked assessment with the transformer framework studied concerning monitoring depression, classifying the level of stress, and detection of suicidal ideation [37]-[40]. The evolving frameworks (e.g. MentalBERT, DeBERTa-v3, ROBERTa-large, etc) are witnessed to excel better performance with a notable average accuracy of 83–89% on multiple datasets. While comparing with existing systems, the proposed study targets towards accomplishing generalizable performance over multiple datasets (CLPsych, EmoReact [<http://multicomp.cs.cmu.edu/resources/emoreact-dataset/>], and DAIC-WOZ), thereby linking psychological theory mapping with behavior classification, which is something being less emphasized in existing approaches.

2.7. Limitations of existing models

Deep learning models, such as BERT and GPT-3, are typically viewed as black-box models, making their predictions difficult to understand. Hence, absence of transparency is one of the potential impediment when such models are used in clinical setting. The conventional methodologies usually encounters issues associated with efficiency and processing time. computational speed is very much essential in clinical setting especially when it comes to carrying out a real-time investigation. The practical usability of the computational model is progressively reduced if it consumes too much duration to yield the

outcome. Usually general dataset are used for training psychological behaviour analysis system that may not be suitable for certain particular disorder in mental health.

The proposed manuscript emphasizes on advancement presented by the ML and NLP towards analysis of psychological behaviour. The notion is towards classifying significant behaviour as well as mapping them to standard benchmarked theories. The identified research gap is bridged by unifying language model with bidirectional encoder along with psychological theory mapping. The overall integrated system results in highly comprehensive framework towards real-time analysis of psychological behavior.

The study model integrates language model with psychological theory mapping which is one of the unique contribution towards investigating mental health. The distinct nature of proposed model can be further contrasted by stating that existing methods usually deploy ML or certain pretrained models which doesn't have much innovativeness nor it offer any breakthrough outcomes. The inference time is quite faster towards behavior classification, unlike existing methods, even compared to a standalone GPT-3 model. The capability of the proposed model for processing voluminous datasets signifies a higher degree of scalability with an effective generalization towards studying various traits of psychological behavior. Its wider applicability, in contrast to any existing models, can be substantiated by its processing and handling capabilities of various types of data, be it from video-based emotional feedback or studying therapy-related transcripts or processing posts from social media. With an extensive analysis of its outcome, the proposed model has proven its efficiency.

2.8. Novelty and Distinction from Related Work

It can be noted that various evolving existing frameworks (e.g. PsychBERT, MentalBERT, MentalHealth-RoBERTa, etc.) are designed on a particular corpus to align themselves with better domain training on mental health specifically. However, these models are eventually meant for emphasizing task-related direct categorization (e.g. detection of suicide ideation, detection of depression, etc.) without proving mapping related to multiple behaviors nor any type of theoretical interpretability. On the contrary, the proposed study model is a way ahead of all these constraints by incorporating an integrated method of mapping that binds established psychological theory with categorized psychological behavior. The key point of novelty contributed by the proposed study is as follows:

- *Mapping Layer with Theory Awareness*: A novel semantic mapping module based on similarity attribute is introduced in the proposed model, connecting formal psychological theory with identified behavior. Hence, the applicability towards clinical settings and interpretability aspect is significantly improved, which is missing in existing models discussed in the literature that are known to facilitate only flat outputs of classifications.
- *Distinct Reasoning with Multi-stage LLM*: The proposed model uses fine-tuned encoder with language model for fusing few-shot and zero-shot prompting. This feature enables to acquisition, categorize, and performance of reasoning about sophisticated behavior with an aid of textual input originating from different sources. This is completely different from existing models that still rely on domain domain-specific transformer.

- *Customized Hybrid Inference*: The presented work encourages semantic alignment and tailored engineering of prompts, unlike conventional models, which demand to be subjected to end-to-end training. This feature facilitates the proposed model to work flexibly on emotional expression, unseen behavior or psychological aspects without any dependencies on retraining.
- *Generalization on Cross-Dataset*: The proposed model has been extensively tested for its effectiveness on five distinct and benchmarked datasets to exhibit its robust performance. On the contrary, the majority of existing models have been evaluated on a single dataset.
- *Mapping Effectiveness Proven with Ablation*: During the analysis of the outcome, an ablation study has been performed to find that classification performance is potentially affected upon removal of psychological theory. This eventually proves that detection performance is not the only sole strength of the proposed model, but it is also capable of semantically grounding the predictive outcomes.

3. Research Method

As a part of extension to our prior study models the fundamental goal of the suggested study model is to create a unique framework for analyzing psychological behaviors by employing LLMs such as bidirectional encoder and language model for behavior categorization and psychological theory mapping. The study's goal is to bridge the gap between NLP and psychological theory by not only categorizing actions (emotions, cognitive distortions, mental states) but also mapping them to existing psychological frameworks. The architecture of the proposed study is shown in Fig. 1

The model performs psychological behavior classification in a novel approach by combining bidirectional encoder and language model. While encoder understands and classifies behavioral features in text, language model generates complex, human-like replies and maps behaviors to psychological theories. This combination takes advantage of both BERT's discriminative power and GPT-3's generative potential, resulting in a powerful and comprehensive framework for psychological study. The specification of LLM models used in the proposed system is as follows:

- *Language model*: An API of OpenAI is used for accessing the GPT-3 model that has been implemented in the proposed study model. The GPT-3.0 model has been implemented considering 156 as the highest number of token length, 40 as the highest k-samples, and 0.7 as temperature for processing tasks related to text generation.
- *Encoder*: The uncased checkpoints for encoder have been deployed in the proposed study model, which are further adjusted for the adopted classification of psychological behavior operation deploying Hugging face transformer library. The study considers 0.00002 as the learning rate on the CLPsych 2015 dataset with a 16 batch size and trained using 5 epochs.

Further, the assessment is also carried out using 5 random seeds associated with each considered model for determining the level of reliability and robustness associated with the outcomes. This is followed by assessing the mean $\pm$ standard deviation associated with the performance metrics of F1-score, recall, precision, accuracy, etc. Further, a statistical paired t-test (or Wilcoxon signed rank test)

was used for better assessment in a pilot study of outcomes. Further bootstrap confidence intervals have been carried out to quantify the ambiguity related to each performance metric. The elaborated discussion of implementing the proposed architecture follows next.

3.1. Data collection and Preprocessing

The process begins with the collection of various high-quality data that accurately describe human psychological behavior in textual form. This information will serve as the basis for training and fine-tuning the model. There are various data sources involved in the proposed study, as exhibited in Fig. 2. The first data source is from therapy transcripts that are verbatim interactions between therapists and patients. Therapy sessions provide valuable information on emotional states, cognitive processes, and behavioral tendencies. The second data source is from social media data is text from sites such as Twitter, Reddit, and forums where people discuss mental health issues, feelings, and personal experiences. The third data source is from survey and interview responses that include organized and unstructured data gathered through surveys and psychological interviews. Participants may share information on their emotional states, behavior, and thought patterns. The fourth data source is from public behavioral datasets, which are data from research initiatives or publicly available datasets on psychological behavior.

After the collection of data is accomplished, preprocessing is carried out. Once acquired, the data is preprocessed to make it suitable for model training. The preprocessing steps include various steps. The primary step is towards cleaning text that involves removing extraneous characters, symbols, and irrelevant material. The next part of implementation is to carry out tokenization that generates a smaller chunks of text (viz. sub-words or words) by breaking the larger segment of the text with a purpose of enhancing the comprehension of the model. The system, further, performs anonymization that consists of targeting and eliminating any form of textual information representing confidential and personal identity from the data in order to comply with ethical standards. At the last step, the model perform normalization that transforms the textual contents into highly consistent structure for minimizing the variance associated with the utilization of the language.

3.2. Selecting Pre-Trained Models

A highly relevant pretrained LLM is chosen by the system for evaluating the psychological behavioral aspect extracted from the textual data. The study opts for language models and bidirectional encoders as the candidate LLM framework for their distinct capabilities. The study opts for LLM model for analyzing as it generates human like textual contents. Apart from this, LLM is quite helpful for assessing open-ended textual information as it facilitates towards extracting the comprehensive context of conversation. It also identifies indicators of emotional text as well as patterns of psychological behavior. Hence, the extraction of logical and deeper insight towards behaviour and emotion can be simplified by adopting LLM in proposed study model. This is quite essential for identifying an unique patterns of psychological traits that is used for realizing even the minor contextual information. Different from GPT, the study opts for bidirectional encoder that considers the contextual information from the pre- and post-

position of target text. Therefore, the complete logical meaning of textual content (in paragraph or sentence) is easily understood. Complicated task like sentiment analysis can be carried out by encoder that consists of manifold behaviour or emotion for classifying the text. All the justified factors lead to choosing LLM in proposed study whose process flow diagram is exhibited in Fig. 3.

The proposed study uses bidirectional encoder and language model towards psychological behavior analysis in the following way:

- *Loading Pre-trained Models:* Bidirectional encoder and language model are both accessible through cloud-based systems or APIs. For instance, encoder can be utilized using libraries such as Hugging Face Transformers, whereas language model is made available via OpenAI's API. In GPT-3, OpenAI provides GPT-3 through an API, allowing the model to comprehend context from the input or produce human-like language. After submitting a prompt to the API, you would get a response that is examined for psychological and emotional information. In the encoder model, a pre-trained model can be loaded using Hugging Face's Transformers library. By including task-specific layers in the model (such as a classifier for emotion detection), encoder can be optimized for psychological tasks like classification.
- *Enriching data quality:* The proposed system performs cleaning and tokenization of text inputs for both models. Preprocessing entails dividing text into smaller chunks (tokens), eliminating stop words, and translating words into model-understandable numerical representations. In this category, language model manages the creation of long-form content; you enter text exactly as it is (such as a treatment transcript), and the model will offer suggestions or extensions. Tokenization and conversion of text into input features, such as token IDs and attention masks, are necessary for a bidirectional encoder.
- *Understanding textual context using LLM:* From the perspective of implementation of language model's generative character, Bidirectional encoder can be asked to compile, finish, or analyze transcripts of treatment sessions, producing possible psychological insights or developing theories regarding the patient's mental health. For instance: "*Can you spot any indications of emotional distress or cognitive distortion in the following therapy transcript?*" bidirectional encoder can be used to categorize behaviors (such as neurotic tendencies or problems with self-esteem) or emotional states (such as happiness, sadness, anger, or anxiety). For instance: "Group the following sentences' emotional tone into categories like joy, sadness, or anxiety."

The prime novelty of this adoption is related to its potential adaptation to psychological behavior. The innovation is in utilizing cutting-edge LLMs, such as bidirectional encoder and language model, to recognize and decipher subtle behavioral, emotional, and cognitive indicators in text. Here, we explicitly want to apply these models to psychological analysis, even though they have traditionally been utilized in general NLP tasks (such as translation, summarization, etc.). Further, the proposed model utilizes the capabilities of both (generation vs. categorization) by combining them to provide a more thorough comprehension of psychological data. While bidirectional encoder's classification strength allows for the

clear characterization of psychological states, language model's generating capacity permits a study of context.

Both bear a potential efficacy towards managing a wider range of tasks related to NLP (detection of emotional state, text classification, etc.), which is one prime reason to adopt in the proposed study model. From the literature, it was noted that certain advanced and evolving approaches like MentalBERT, DeBERTa-v3, and RoBERTa-large have also been reported to offer better performance on NLP tasks; however, they are not included in the proposed study due to the following reasons:

- Both DeBERTa-v3 and RoBERTa-large have been proven to be effective for analyzing a large volume of NLP operations; however, their performance is not up to the mark when compared with standalone BERT related to the analysis of particular psychological behavior.
- For studying applications of mental health, the MentalBERT is also reported to offer better outcomes as they are tailored version towards domain-specific models. However, it is noted that when standalone bidirectional encoder is fused with psychological theory mapping, the performance is quite progressive better with better interpretability. It assists in acquiring emotional states and cognitive distortions, which MentalBERT is found to offer with sub-optimal outcomes.

Therefore, the model development of the proposed study doesn't include the above-mentioned LLM models as there is a higher likelihood of inducing extra computational complexity leading to increased duration of generating inference. Considering the adopted use case in the proposed study, the adoption of LLM is less likely to contribute towards either better interpretability or enhanced model performance. On the contrary, the proposed study targets introducing a method that offers a fair balance between interpretability and efficiency for varied demands towards clinical psychology.

3.3. Model Finetuning

Following their selection, the pre-trained models need to be adjusted for the particular psychological behavior analysis task. The act of tailoring a general-purpose pre-trained model to a specialized domain—in this example, psychology—is known as fine-tuning. Although the base models are good at interpreting language, psychological analysis is not their primary function, therefore, this stage is essential.

- *Collecting and Preparing Domain-Specific Data:* To perform model finetuning, the proposed system uses labelled examples of psychological activity (e.g., emotional states, cognitive biases, mental health symptoms). For this purpose, a dataset is considered of therapy session dialogues annotated with emotional states (anxiety, despair, happiness), cognitive distortions (overgeneralization, black-and-white thinking), and behavior patterns (avoidance, impulsivity). Further, text preprocessing is carried out by text tokenization (breaking down sentences into tokens), padding (for consistency), and confirming text compatibility with the model's input format are all part of preprocessing the data.

- *Bidirectional encoder finetuning*: For classification tasks such as emotion or behavior detection, a task-specific output layer (such as a SoftMax layer for multi-class classification) is added to the pre-trained bidirectional encoder model. The model is trained using the domain-specific dataset. During fine-tuning, the model learns to recognize patterns in the text that indicate psychological activities, such as identifying specific phrases or language markers that indicate worry, self-doubt, or negative thought patterns. For instance, if a line like "*I feel like everything is going wrong and nothing will improve*," bidirectional encoder will recognize it as a cognitive distortion, such as catastrophizing. To maximize performance, the learning rate, batch size, epoch count, and other parameters are modified. The study uses random search in order to explore the optimal hyperparameters for this particular study.
- *Adjusting Language Model*: It is necessary to know that training of the language models are not carried out by supervised mechanism always. However, prompt engineering is opted for adjusting the model by creating an accurate prompt for exhibiting an essential insights of psychological behaviour. It is to be noted that labelled dataset (BERT) is not always adjusted by language model, but there are also higher side feasibility of using unified human feedback with reinforcement learning for performing adjustment that is highly supported by OpenAI. Adoption of such mechanism can be useful towards enhancing the accuracy of the language model in identifying the psychological behavior. It is done by facilitating input (human-like) throughout the complete process of training.
- *Validation of the framework*: The proposed system ensures a mandatory testing for both bidirectional encoder and the language model adopting cross-validation method. This is done in order to resist any event of overfitting and ensure better generalization of the computational framework to prior psychological data that is unreported. After the adjustment is carried out, the performance of the models are further assessed adopting standard metrics of F1-score, recall, precision, accuracy, etc.

The innovativeness of this part study implementation is mainly linked with the customizable and behavior model with enhanced contextual adaptation. The proposed model is capable of customizing a highly generalized model to a particular field associated with analysis of psychological behaviour. One of the innovative usage resides in identification of particular linguistic pattern in therapeutic conversation that the proposed system achieves by adjusting bidirectional encoder. Apart from this, the study model can also alter its responses based on task while an accurate classification is facilitated by the supervised adjustment carried out by the bidirectional encoder.

3.4. Classification of Behavior

The proposed computational framework contributes towards classifying and categorize social media post and transcripts of therapy session that are basically in textual form into types of psychological behaviour. The module is meant for determining different forms of psychological indicators, cognitive distortions, emotions, patterns etc. in order to perform classification. It is an essential step of operation

as it facilitates the system to categorize behavioral features, mental health disorders, cognitive distortions, and emotional states. The internal steps of this process is briefed as follows:

- *Define Essential Categories:* The preliminary step is towards determining the form of psychological behaviour towards assisting in classification. The determination of these categories are carried out by the agenda formulated for the system e.g. i) frequently witnessed patterns of behavior, ii) various types of conditions associated with manifold mental disorder, iii) different types of personality, iv) cognitive distortions, and v) emotional states.
- *Prepare and Label the Dataset:* The next stage is to collect and prepare a dataset that will be used to train and validate the model. This dataset must include labelled examples of psychological activity, including text linked with each category. For supervised learning, data must be labeled manually or semi-automatically. For example, A remark like "*I feel like everything is falling apart, and I can't do anything right*" in a therapy session transcript may indicate anxiety or cognitive distortion (catastrophizing). A social media post stating, "I never seem to fit in anywhere," may indicate melancholy or low self-esteem.
- *Textual Data Preprocessing:* Preprocessing is a necessary step before entering data into the model for behavior classification. Preprocessing cleans and transforms text into a format that ML models can understand. Typical preprocessing procedures include normalization, tokenization, stopword removal, stemming, and padding.
- *Model Selection for Behavior Classification:* The proposed study uses bidirectional encoder for classification problems in this framework because it is bidirectional and can grasp the context around words (both before and after) in a phrase. Bidirectional encoder is especially beneficial for tasks that require understanding the context of words and phrases in relation to psychological activity. Model Architecture of bidirectional encoder is as follows: i) a pre-trained bidirectional encoder model (e.g., encoder-base, encoder-large) is used and trained on large general-domain corpora, ii) Tokenization is carried out of input text results in input IDs, attention masks, and token type IDs for the bidirectional encoder model to process, iii) Within the classification layer, after the pre-trained bidirectional encoder layers, a custom classification head is applied. This classification head includes a fully connected layer that maps the bidirectional encoder output to a probability distribution across classes (e.g., emotional states, cognitive distortions). Finally, a SoftMax layer is used to normalize the output into probabilities that add up to one.
- *Training the Behavior Classification Model:* The next step after defining the model architecture is to train the model on the labeled dataset. The training method is putting preprocessed text input into the model and optimizing the weights to minimize a loss function, usually cross-entropy loss, for classification tasks. The process involved in training are i) Loss Function where during training, the model predicts the psychological behavior label for each input text and uses the cross-entropy loss function to calculate the error between predicted and true labels, ii) Backpropagation where the error propagates back through the network, adjusting the model weights to minimize loss, iii) usage of an optimization algorithm where the Adam optimizer adjusts weights and reduces the loss function across numerous iterations (epochs), and iv) usage of early stopping where this strategy

stops training when the model's performance on a validation set does not increase after multiple epochs, preventing overfitting.

- *Evaluation of Model:* The assessment of the model has been carried out after the training in order to find out its performance on priorly dynamic data. Similar accuracy-based standard metrics have been used for assessing the ability of the model.
- *Post-Processing of an Outcome:* Once the classification task towards psychological behaviour has been accomplished by the model, the post processing operation is carried out for the outcome in order to enhance its usability. For this purpose, the raw predictive value of the model is converted to interpretable information. A categorization report is preliminarily created with certain information related to prediction of a model and performance of class behavior. A probability distribution is generated by the framework towards each categories. One example to understand this is as follows: consider a response “*I am so overwhelmed that everything looks like incorrect*” obtained from transcript of a therapy. The outcome of analysis of this text will look like 85% of anxiety. Such issues towards classification can be evaded by this model where the system links an application with the presented model of behavioral classification in real-time. The module contributes towards depiction of cognitive and emotional states identified within the text obtained from the recommendation and insights.

The proposed study model contributes towards a novel and efficient classification model where a pretrained language model is chosen by the system followed by adjusting. The model assists toward determining the states of an emotion, distortions in cognitive traits, and different behaviour of psychological textual data. A task-oriented mechanism has been introduced by the proposed system that emphasizes on particular classes. The framework is customized towards different types of application e.g., performing sentiment analysis of social media post, monitoring mental health, and analysis of therapy session.

3.5. Psychological Theory Mapping

The proposed study uses Psychological Theory Mapping that connects to the classification of a model with conventionally existing ideas of psychology. There are various standards that define such theories to be a part of clinical framework (viz. updated criteria of diagnosis), cognitive (Beck's Cognitive Triad), and emotional (James-Lange theory of emotion). The proposed study generates a context towards the raw and original data by connecting the standard psychological theory with behavioral classifications.

- *Define Psychological Theories for Mapping:* The primary task towards mapping the above-mentioned theory is to understand its insight. From the perspective of emotion theories, there are various models, viz. Schachter-Singer Theory [44], Cannon-Bard Theory [45], and James-Lange Theory of Emotion [46]. From the perspective of cognitive theories, there are various models, viz. Beck's Cognitive Triad is a model linked to depression that highlights negative thinking about oneself, the world, and the future. Cognitive distortions include catastrophizing, overgeneralization, and black-and-white thinking. From the perspective of personality theories, the Big Five personality traits

include openness, conscientiousness, extraversion, agreeableness, and neuroticism, while Freudian psychoanalysis examines unconscious drives, including defense strategies. The model also supports the usage of the DSM-5 standard for the mental health model that is used for analyzing various mental disorders [47].

- *Create a Mapping Framework:* Once the psychological theories have been identified, we may design a mapping framework (Table 1) that connects the behavior categorization model's output classes (such as emotional states or cognitive distortions) to specific psychological theories. This framework serves as a reference for interpreting the model's results in terms of psychological constructs.
- *Integrate the Behavior Classifier with Psychological Theory:* Once we have the mapping structure, the next step is to incorporate it into the system. The behavior classification model (for example, bidirectional encoder for classification tasks) produces outputs (such as emotional state labels), which we will interpret using the psychological theory mapping. The initiation of the process is carried out using preprocessed text that is considered as an input to the behaviour classification model that yields a predicted labels (e.g. depression, catastrophizing, anxiety, etc.). Depending upon this outcome, the proposed system utilizes a mapping model for matching the anticipated labels to a suitable psychological theory. Suppose the prediction of the model leads to yield anxiety than it can associated with standard theory of Cannon-Bard or James-Lange that eventually defines anxiety. After linking the classified outcome to the theory, the consecutive operation is to carry out result interpretation in terms of psychology. For an example, outcome with inferior self-esteem will be equivalent to Beck's Cognitive Triad where the pattern of behaviour is equivalent to negative self-evaluation. At the end, the outcome is mapped to the psychological theory to generate recommendations associated with the experts of mental health. For an example, if the system detects anxiety than it will offers ways for regulating it depending upon the theories of emotion regulation such as cognitive behavioral therapy to be used for controlling anxiety.
- *Feedback Personalization using Mapping:* The proposed scheme adopts psychological theory mapping for yielding highly customized feedback depending upon textual data. Such form of customized outcome in form of feedback can be utilized for maximizing self-awareness, recommending intervention, and leading sessions of treatments. Depending upon the applicable theory, a distinct psychological insights are obtained by the proposed system. It will mean that if the outcome suggested reduced self-esteem than the feedback system will involve recommendation based on CBT that is meant for changing the negative thought patterns. Not only this, the proposed study model can also offer a very specific recommendation towards behavioral or cognitive methods. In such case, the recommendation system may involve suggesting specific task (e.g. though record sheet or activities related to journalling) when CBT is suggested for discovered symptoms of depression. Apart from this, a therapeutic advice can be generated by the system depending upon the psychological theories towards therapy-based applications. For an example, the system can specifically recommend various discreet intervention e.g. impulse control exercises or mindfulness practices when identified with ADHD or anything linked to impulsivity.

Table 1
Essential elements of the mapping framework

LLM Behaviour Classification	Psychological Theory	Psychological Construct
Impulsivity	Freud's Psychoanalysis (Id-driven)	Behavioral Pattern (Lack of Self-Control)
Depression	DSM-5 Criteria, Cognitive Triad	Mental Health Disorder (Depressive Episode)
Low Self-Esteem	Beck's Cognitive Triad	Negative Self-Perception
Catastrophizing	Beck's Cognitive Distortions	Cognitive Bias (Negative Thinking Patterns)
Anxiety	James-Lange Theory, Cannon-Bard Theory	Emotional Response to Stress or Threat

The key contribution of this operational module is towards efficient contextualization as well as interpretation of the original behaviour that is further justified by standard theories and frameworks of psychology. On the basis of informed psychological theory, the proposed system facilitates practical insights thereby enhancing the quality of feedback on the basis of the theory mapping. This operation leads to enhanced specialized case of mental health that is practically more viable. The adopted methodology offers an evidence that significance of model outcome and behavioral classification are quite higher even extending to practical solution towards enhancing mental health.

3.6. Ethics Statement

All methods were carried out in accordance with relevant guidelines and regulations. This study utilizes secondary datasets, namely the CLPsych 2015 dataset and the DAIC-WOZ database, which contain anonymized data derived from human participants. The original data collection processes for these datasets were conducted under appropriate ethical approvals from relevant institutional review boards (IRBs) or ethics committees. Informed consent was obtained from all participants involved in the original studies. As this work involves only the analysis of previously collected and anonymized data, and does not include direct interaction with human subjects, additional ethical approval was not required. All analyses were conducted in compliance with established ethical standards for research involving human-derived data. Our work does not involve direct interaction with human subjects. Further, we have complied with the terms of usage of these dataset as well as obeyed the privacy policies updated by the curators of original dataset. The next section illustrates the outcome accomplished from this study's implementation.

4. Result Analysis

The simulation parameters determine the overall environment in which the models are trained and evaluated. The parameters include computing resource selection, data splitting algorithms, and

evaluation protocols. The dataset for this study comprised 50,000 labelled instances representing various psychological behaviors. The data was separated as follows: i) The model was trained on 80% of the training set (40,000 instances), ii) Validation Set: 10% (5,000 instances) were utilized to modify the model's hyperparameters. iii) A test set of 10% (5,000 instances) is employed to evaluate model performance.

It should be noted that the assessment being carried out for the proposed system has used a specific version of software and hardware in order to test them in real-time settings. The adopted settings have contributed to approximately 160ms per instance for observed inference time. The implementation has been carried out on multiple frameworks, viz. CUDA 11.8, Transformers (HuggingFace), and PyTorch 2.1 considering cloud infrastructure of AWS EC2 with p4d instances in order to accomplish low latency with Elastic Fabric Adapter. The batch size is retained at 16 during training operation, while testing for latency during inference is carried out with single instances. The system uses 256 GB DD4 RAM as memory on a 32-core Intel Xeon Gold 6338 CPU and NVIDIA A100 40 GB as GPU for both the LLM models layers. Table 2 showcases the latency breakdown, which is average per instance.

Table 2
Latency breakdown

Component	Time (ms)
Post-processing & formatting	10
Mapping logic (embedding + similarity computation)	15
GPT-3 API call (when used)	50
Bidirectional Encoder Encoding (forward pass)	80
Tokenization (HuggingFace)	15
Total	~ 170

The few-shot classification is carried out selectively using API calls of language model followed by further deploying prompt engineering and caching to enhance its performance. It is also noted that average latency reduces to approximately 140ms if only a fine-tuned bidirectional encoder model is used. The average latency of response is found to be 50ms, which includes both prompt transmission and token serialization for integration of API in case of language model. The cumulative latency discussed in Table 2 consists of all components from the start to the end of the overall working principle of the proposed study model. The next section outlines the adopted dataset.

4.1. Dataset Adoption

This study used a combination of publicly accessible and bespoke datasets to train and evaluate the suggested framework for psychological behavior analysis. These datasets supplied a wide range of textual data reflecting psychological actions and emotional states, which was necessary for developing

a strong model capable of classifying and mapping behaviors to recognized psychological theories. The datasets were selected to capture a wide range of psychological illnesses, such as depression, anxiety, stress, and general emotional states, ensuring the model's versatility and generalizability.

The dataset that has been used for experimentation are as follows:

- *DAIC-WOZ (Dialogue Act Corpus for Depression Detection)* [36]: This is the most frequently adopted dataset deployed towards investigating the depression-based indicators using NLP. The dataset consists of 2000 + transcripts of audio and text related to clinical procedure undertaken by clinicians. The textual data consists of exchanges of conversation towards representing various states of mental health viz. emotional discomfort, anxiety, despair, etc. This is an ideal dataset for identifying a smaller changes in the signals within the conversational text. The severity of depression is indicated by the labels permitting the framework for learning both emotional intensity and behaviour classification.
- *The CLPsych 2015 dataset* [48]: This is another unique dataset that consists of emotional state indicators derived from textual contents shared within social post. It consist of 1000 + labelled contents (post) representing mental state of users sharing their views and opinions in social media. The CLPsych dataset is useful because it contains text data from a variety of sources, including blogs and online forums, which show a greater range of psychological behaviors across persons and emotional states.
- *EmoReact Dataset* [49]: The EmoReact collection includes brief video scripts and annotations of emotional reactions like rage, fear, happiness, and melancholy. Although originally intended for emotional state detection, it is useful for training models to recognize emotional expressions in text. While not solely concerned with mental health, this dataset is beneficial for recognizing high-level emotions and linking them to psychological theories such as the Big Five personality traits and cognitive biases in psychological assessments.

Apart from this, the proposed study has also considered custom therapy data and custom social media data. All the above-mentioned dataset retains proper anonymization while being publicly available. A standard consideration of 80% share of data to be used as a training, 10% share of data to be used as a validation set, while the remaining 10% share of data to be used for testing has been considered in this assessment for each dataset. The experts in psychology performed an annotation task while labelling each instance, which was also carried out in accordance with the psychological model. Several researchers have extensively investigated behavior analysis in the past to extract meaningful information [50–57].

4.2. Adopted Hyperparameters

The adoption of hyperparameters is discussed as follows:

- *Hyperparameters for Bidirectional Encode Model*: To fine-tune the bidirectional encoder model, the proposed study uses various hyperparameters. Adam optimizer has a learning rate of 0.00002 (with

decay). This learning rate was chosen to strike a compromise between convergence speed and stability throughout training. A batch size = 16 is used to fit the model into the available hardware while maintaining a reliable gradient descent process. The epoch count = 4 is used to avoid overfitting and ensure adequate data exposure while the model was trained over four epochs. The warm-up steps = 500, which are the steps in the early training phase where the learning rate is gradually increased to avoid abrupt gradient changes. Finally, weight decay = 0.01 is used to penalize big weights and prevent the model from overfitting. All these are general hyperparameters, while the proposed system is also characterized by model-specific hyperparameters which are as follows: A maximum sequence length of 128 Tokens is considered to set the maximum length of the input sequences. Sentences more than this length are trimmed, while shorter sentences are padded to ensure consistency. The hidden size = 768 is considered as the amount of hidden units in each layer of the model determines its ability to learn complicated representations of text. The attention heads = 12 is considered where the transformer architecture's multi-head attention mechanism employs a specific number of attention heads. Finally, Layer count = 12 is considered as bidirectional encode's architecture consists of 12 transformer layers that aid in the creation of representations from input sequences.

- *Hyperparameters for Language Model:* For language mode, we used the OpenAI API (since language mode is not directly fine-tuned but is accessed through pre-trained versions). However, we changed specific parameters to fit language mode with the aim of psychological behavior classification: The temperature is set to 0.7, which determines the randomness of the model's outputs. The creativity is balanced with the determinism by adoption 0.7 as an empirical score of temperature for providing higher degree of consistency in the answers while there are present higher diversification towards carrying out predictive analysis. The study considers 200 as maximum set of token for indicating chunks of words generated for each answer leading to logically enhanced information as an outcome. The assessment is carried out considering 0.9 as value of nucleus sampling Top_p for limiting the outcome of framework to the tokens with higher feasibility while the total score of probability is retained more than 0.9. The study uses prompt engineering towards fine-tuning where API based language model is used for generating outcomes with enriched contextual information. Apart from this, the study model also develops different types of prompts for identifying behaviour, cognitive distortion, emotion, etc utilizing textual data. Zero-shot learning has been used by the system while for specific behaviors (such as depression identification), a few-shot learning strategy was applied, in which a few samples of annotated behavior data were provided to assist the model in understanding the context.
- *Hyperparameters for Psychological Theory Mapping:* Similarity threshold is set to 0.85. When mapping classifications to psychological theories, output behavior was assigned to a theory if the semantic similarity between the behavior label and the theory exceeded 85%. This guaranteed that the theory mapping was accurate and context-appropriate. To prevent overfitting, each psychological theory was limited to 50 mapping categories. This ensured that only the most relevant theories were used when mapping behavioral categorization. The confidence level for mapping is 0.8. When the classification model output had a confidence level higher than 80%, it was mapped to

the associated psychological theory. This ensured that less certain classifications were not hastily mapped. Table 3 showcases the hyperparameters used.

Table 3
Summarized Hyperparameters Adoption

Parameter	Bidirectional encoder	Language model
Learning Rate	0.00002	N/A
Batch Size	16	N/A (API-based)
Epochs	4	N/A (Pre-trained API)
Max Sequence Length	128	200 tokens
Hidden Size	768	N/A (API-based)
Temperature	N/A	0.7
Top-p (nucleus sampling)	N/A	0.9
Mapping Accuracy Threshold	N/A	85% similarity

The search towards effective hyperparameters has been carried out for both the LLM models. Specifically for bidirectional encoder, the involved test parameters are learning rates of 0.00001, 0.00002, and 0.00003. The batch sizes considered are 8, 16, and 32, while the Adam optimizer has been used with a weight decay of 0.01, and dropout rates of 0.1 and 0.2 have been used. For language model, the parameters involved are (0.7, 1.0) as temperature, (256, 512) as maximum length of token, and (20, 40, 60) as highest *k*-sampling. The optimal configurations were selected depending on the performance of the validation set, where bidirectional encoder recorded its optimal performance at 0.00002 learning rate, while language model has recorded its optimal performance at 0.7 temperature.

4.3. Accomplished Study Outcome

This section discusses in detail the results gained from the suggested framework for psychological behavior analysis using Large Language Models (LLMs) such as bidirectional encoder and language model. Various accuracy-based performance parameters as well as classification time has been used to assess the proposed model as well as existing methods. According to the outcomes of Table 4, it can be seen that proposed system excels better performance in contrast to state-of-the-art methods. The primary performance metric of accuracy is responsible for evaluating the ability of the model to categorize behaviour with higher degree of correctness. Various forms of behaviors that are identified are different psychological activities, cognitive distortion, emotional states, etc. From the outcome of the Table 4, it can be noted that highest score of accuracy has been scored by the proposed system (*acc* = 87.5%) in contrast to three different State-of-the-Art (SoA) methods. The second metric of precision is used for evaluating the segment of true positive outcomes divided by all the positive instances. While the outcomes of precision metric among SoA methods ranges between 79.3–84.5%, the proposed system scored highest precision of 88.2%. The outcome states that proposed solution can be stated as

dependable solution when it comes to predict a specific behaviour or psychological state. Similar trend is also noted for the recall score of proposed system which shows 86.9% that exhibits potential capacity to detect potential psychological activities within the dataset. The system then computes the harmonic mean of precision and recall in order to assess the trends of F1-score. The outcome shows 87.5% of higher F1-score accomplished by proposed study model which states a fair degree of equilibrium between recall and precision. Further, it is noted that textual data processing is quite efficient and faster when subjected to proposed model which consumes only 1.2 seconds for processing each instance. This is much faster than SoA methods which consumes duration ranging between 1.8–2.5 seconds. The last performance metric is mapping accuracy which is responsible for evaluating the degree of success that the system matches with the classified behaviour to specific psychological theory.

Table 4
Study Outcome

Performance Metric	Proposed System	Method A (SoA) [13]-[15]	Method B (SoA) [16]-[27]	Method C (SoA) [38]-[31]
Accuracy (%)	87.5	82.3	80.1	78.5
Precision (%)	88.2	84.5	81.7	79.3
Recall (%)	86.9	81.0	79.5	76.8
F1-Score (%)	87.5	82.7	80.6	78.1
Classification Time (s)	1.2	1.8	2.1	2.5
Mapping Accuracy (%)	90.0	83.2	80.0	75.4

The proposed system has a mapping accuracy of 90%, exceeding all SoA approaches by a significant margin, ensuring that the classifications are not only accurate but also interpretable and relevant in terms of psychological theory. Apart from this, an extensive analysis is carried out as shown in Table 5. The outcome of the study has been assessed by comparing it with various existing methods deployed for analysis of psychological behavior (e.g., CNN, Random Forest, SVM). Although these approaches have been reported for their effectiveness, the proposed system surpasses the overall performance considering every aspect of performance metrics. The accuracy performance is quite high for the proposed system for all the datasets. Further, consistency is witnessed for F1-score, recall, and precision, exhibiting the capability of the proposed system to identify and precisely categorize behavior of an individual. The performance of mapping accuracy is also witnessed to be 90% for proposed system, which is significantly higher than state-of-the-art methods, whose score ranges between 75–83%. It is also noted that the capability of Random Forest and SVM is restricted towards acquiring the complex form of linguistic patterns, mainly due to their dependencies on features. On the contrary, the adoption of bidirectional encoder and language model as LLM models in the proposed study facilitates an effective contextual text understanding for enhanced ability towards classification. The challenges

towards accomplishing an interpretable outcome are quite high for existing methods where the classification results have been presented without any background of contextual support. On the contrary, proposed model offers a significant interpretable predictive outcome, which is mainly due to the integration of psychological theory mapping. This facilitates a ground truth reference stating proper alignment of classification outcome with standard concepts of psychology.

Table 5
Extensive outcome of study analysis

Dataset	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)	Inference Time (ms)
DAIC-WOZ	Proposed System	91.5	89.3	92.1	90.7	94.8	150
	SVM	85.3	83.5	85.1	84.3	89.2	200
	Random Forest	87.2	85.7	87.5	86.6	91.4	180
CLPsych 2015	Proposed System	88.2	85.9	89.7	87.7	92.5	140
	BERT	82.5	80.3	81.9	81.1	87.4	250
	GPT-3	83.9	82.4	84.1	83.2	88.6	300
EmoReact	Proposed System	93.1	91.6	94.2	92.9	96.2	160
	CNN	85.7	82.8	84.5	83.6	89.8	210
	RNN	86.4	84.1	86.2	85.1	90.4	190
Custom Therapy Data	Proposed System	92.0	90.4	93.0	91.7	95.5	145
	SVM	84.8	82.1	84.0	83.0	88.9	210
	Random Forest	86.1	84.5	85.8	85.1	90.3	190
Custom Social Media	Proposed System	90.3	88.7	91.2	89.9	93.7	170
	BERT	83.6	81.5	82.8	82.1	87.5	250
	GPT-3	85.2	83.4	85.6	84.5	88.9	280

Finally, the inference time for the proposed system is found to be quite faster (150-170ms) when it is compared with that of other existing methods (250-300ms). These outcomes exhibit the better applicability of the proposed model in practical world applications. Table 6 showcases the computation of mean $\pm$ SD (standard deviation) considering multiple random seeds over 5 runs. The analysis is

carried out considering the Wilcoxon Signed-rank test (paired t-test) in order to validate the statistical significance for each model pair. The outcome shows a highly statistically significant difference with $p < 0.05$ for the proposed and baseline models. The outcome of Table 5 showcases the robustness of the proposed system.

Table 6
Analysis of mean and standard deviation

Metric	Proposed System	Method A (SoA) [13]-[15]	Method B (SoA) [16]-[27]	Method C (SoA) [38]-[31]
Accuracy (%)	87.5 ± 1.2	82.3 ± 1.5	80.1 ± 1.0	78.5 ± 1.3
Precision (%)	88.2 ± 1.1	84.5 ± 1.3	81.7 ± 1.2	79.3 ± 1.4
Recall (%)	86.9 ± 1.3	81.0 ± 1.2	79.5 ± 1.1	76.8 ± 1.2
F1-Score (%)	87.5 ± 1.1	82.7 ± 1.4	80.6 ± 1.2	78.1 ± 1.3
Inference Time (s)	1.2 ± 0.1	1.8 ± 0.2	2.1 ± 0.3	2.5 ± 0.2

Apart from the above-mentioned Table 6, an extended comparative analysis has been carried out between proposed model and conventional benchmarked models as well as varied form of methodologies reported in literatures towards analysis of psychological behaviour (Table 7). From the perspective of comparison with literatures, the studies carried out by Beck Depression Inventory [14], Udding et al. [16], Kerasiotis et al. [17], Reitberg et al. [24], Montejo-Ráez et al. [4], Hussain et al. [19], Zhang et al. [18], MentalHealth-RoBERTa [32], PsyQA-BERT [33], Malgaroli et al. [15], and Salas-Zárate et al. [2] have been chosen.

Table 7
Comparison of Proposed Framework with Existing Studies

Existing Studies	Model	Dataset Used	Performance Metrics	Key Outcome
Salas-Zárate et al. [2]	Bi-LSTM	Twitter User Comments	Acc: 86.5% F1: 85.3%	-Identifies levels of depression -Lacks interpretability
Montejo-Ráez et al. [4]	RoBERTa-Large	Twitter Mental Health Corpus	Acc: 87.3% F1: 86.7%	-Enrich with Context awareness -lacks psychological linkage
Beck Depression Inventory [14]	Logistic Regression	BDI Clinical Dataset	Acc: 79.5% F1: 78.2%	-Early standardized approach; -No contextual text analysis
Malgaroli et al. [15]	Hybrid CNN-LSTM	Therapy Session Corpus	Acc: 87.9% F1: 86.6%	-Good contextual extraction -Sub-optimal generalization
Udding et al. [16]	RNN	Reddit Depression Forum	Acc: 82.6% F1: 81.3%	-Highly effective for identifying emotional tone -Do not possess potential theoretical grounding
Kerasiotis et al. [17]	CNN	Social Media Posts	Acc: 83.4% F1: 82.5%	-Optimal performance on surface-level emotion cues
Zhang et al. [18]	XLNet	CLPsych 2021	Acc: 86.2% F1: 87.3%	-Acquires long-range context -No mapping of theory
Hussain et al. [19]	DistilBERT	Mental Health Tweets Dataset	Acc: 84.1% F1: 83.8%	-Lightweight transformer design; -Moderate interpretability

Existing Studies	Model	Dataset Used	Performance Metrics	Key Outcome
Reitberg et al. [24]	BERT	Clinical Text Corpus	Acc: 83.2% F1: 83.1%	-Potential linguistic context -Restricted interpretability
MentalHealth-RoBERTa [32]	Fine-tuned Transformer	Reddit Health Forums	Acc: 86.2% F1: 85.6%	-Domain-optimized performance -Study restricted to suicidal ideation only
PsyQA-BERT [33]	CBT-Aligned BERT	PsyQA Dataset	Acc: 87.4% F1: 86.8%	-Theory-guided design -Lacks deployment in real-time
SVM (Baseline)	SVM	CLPsych 2015	Acc: 84.2% F1: 83.2%	-High precision -Minimal generalization
Random Forest (Baseline)	RF	DAIC-WOZ	Acc: 86.1% F1: 85.5%	-Robust performance -Slow inference
GPT-3 (Standalone)	GPT-3	Custom Therapy Data	Acc: 84.1% F1: 83.4%	-Yield contextual responses, -Limited stable classification
Proposed Model	Hybrid LLM Framework	DAIC-WOZ, EmoReact, CLPsych 2015, Custom Data	Acc: 91.5–93.1% F1: 90.7–92.9% Mapping Acc: 90%	-Maximized interpretability, -Enhanced context linkage, -Supports real-time inference (~ 150 ms)

The rationale for opting them are as follows: The Beck Depression Inventory [14] has been stated as traditional benchmarks towards both manual and questionnaire-based evaluation process. The selection of Udding et al. [16] and Kerasiotis et al. [17] is done owing to their pioneering deployment of RNN and CNN towards task related to recognition of an emotion. The work of Montejo-Raez et al.[4] and Reitberg et al. [24] have been selected as they claims of successful implications of BERT and RoBERTa that are transformer models used for understanding clinical text. Zhang et al. [18] and Hussain et al. [19] study model has been selected as they introduce XLNet and DistilBERT that are much contextually enriched

and lighter variants for ensuring handling of long-range dependencies and computational efficiency. MentalHealth-RoBERTa [32] and PsyQA-BERT [33] are another benchmarked transformer models that are curated specifically towards the context of mental health. Finally, work done by Salas-Zarate et al. [2] and Malgaroli et al. [15] has presented deployment of NLP and hybrid deep learning models for combining social media data and multimodal therapy. Apart from these existing benchmarked approaches, different SoA methods e.g. Random Forest, SVM and transformer-oriented approaches of GPT-3.

The outcome of Table 7 clear showcases that performance of proposed model has excelled significantly better in contrast to both baseline models as well as existing SoA methods discussed in literatures. Table 7 also showcases that proposed model advances very clearly over both conventional and transformer models. Various conventional models of ML (SVM, Random Forest) is found to exhibit highly restricted contextual sensitivity while RoBERTa, XLNet, encoder etc. being a transformer model is found to showcase enhanced understanding capability of text. However, these models lacks interpretability quite significantly. Hence, the study outcome showcase highly optimal and efficient performance of proposed model that is noticeably recorded to possess better outcomes.

4.4. Ablation Study

An ablation study is carried out to assess the overall performance due to the influence of the psychological theory mapping. From Table 6 and Table 7, it is noted that the model completely depends upon the outcome of bidirectional encoder and language model during mapping-disabled configurations for classifying behavior where there is absolutely no mapping to any apriori theory of psychology. An accuracy of 83.2%, precision of 84.5%, recall of 82.1%, and F1-score of 83.3% have been witnessed for the ablation station with mapping disabled. During mapping-enabled configuration, the study accomplishes 87.5% accuracy, 88.2% precision, 86.9% recall, and 87.5% of F1-score. There is a notable difference in the outcomes, stating an important role played by psychological theory mapping towards system interpretability.

Table 8
Outcome of ablation study

Metric	Proposed System (with Mapping)	Mapping Disabled System
Accuracy (%)	87.5	83.2
Precision (%)	88.2	84.5
Recall (%)	86.9	82.1
F1-Score (%)	87.5	83.3

The outcome exhibited in Table 8 showcases potential performance degradation when the mapping components are disabled. The system witnesses an 83.2% accuracy drop from 87.5% accuracy in the ablation study. This outcome itself exhibits the significance of mapping steps in order to map the classification of behavior to the standard theory of psychology for better accuracy and interpretability.

Further, an extended ablation study has been carried out whose outcome is presented in Table 9. For this purpose six configurations C1-C6 has been considered for isolating the contribution linked to encoder, language model, and PTM module involved in proposed study. The outcome showcases are means over the test scores associated with the benchmarked datasets adopted in study. The mean score of dataset is considered as metric values while the inference time is considered as median scores for each instances of latencies evaluated on deployed model. Wilcoxon signed-rank test is performed as a statistical measurement to exhibit that proposed model (C6) showcases superior performance in ablation study with $p < 0.05$ as it outperforms optimal ablated alternatives.

Table 9 outlines the quantitative ablation results associated with proposed framework towards analysis of psychological behaviour. The different types of configurations are either disabled selectively or isolation is carried out on core components e.g. PTM, language model, and encoder for quantifying their combined as well as individual effects. The rationale of opting each configurations are as follows: Configurations C1 and C2 represents the baseline for all single architecture while configuration C3 represents impact of ensemble architecture without any PTM. Configuration C4 and C5 is meant for quantifying contribution of PTM when subjected to pairing with only one backbone. Finally, configuration C6 represents complete synergy of system followed by justifying the complexities.

The outcome showcases that maximized overall accuracy (91.5%) is accomplished by full configuration (i.e. C6) while highly balanced score of precision and recall is recorded with 91.6% and 91.3% while real-time inference of approximately 155 ms has been witnessed in ablation study. The enhancement of 4–6% approximately is recorded over ablated variants while this outcomes states that proposed model substantially improvise the robustness as well as interpretability without any adverse effect computational efficiency.

Table 9
Extended Ablation Study

Config	Short name	Feature extraction	Model composition	PTM (mapping)	Accuracy (%)	Inference time (ms)
C1	BERT-only	BERT contextual embeddings (CLS token + pooled features)	Fine-tuned BERT classifier (task head)	No	85.9	151
C2	GPT-3-only	Prompt-based GPT-3 semantic outputs (few-shot prompts → embedding/prob)	GPT-3 used as direct classifier via prompt templates (API)	No	86.3	291
C3	BERT + GPT-3 (no PTM)	BERT embeddings + GPT-3 prompt outputs fused	Late fusion ensemble: weighted average of BERT and GPT-3 prediction scores	No	87.9	211
C4	BERT + PTM	BERT embeddings + semantic similarity vectors to theory lexicon	BERT classifier + PTM post-processor (mapping reranks/pseudo-features)	Yes	89.6	171
C5	GPT-3 + PTM	GPT-3 prompt outputs + mapping features	GPT-3 classifier + PTM post-processor (mapping + reranking) — selective GPT-3 calls via cache	Yes	92.2	181
C6	Proposed	Fused (context + generative reasoning) + mapping vectors	Hybrid: fine-tuned BERT + selective GPT-3 reasoning + PTM for final reranking & explanation	Yes	92.6	166

According to the outcomes of Table 9, the configuration C1 isolates baseline transformer classification while it measures encoder alone. C2 using language model is witnessed to isolate baselines that are generator centric and can measure classification capability of language model along with certain overhead of API. C3 is noted to measure synergy of discriminative as well as generative signals without any dependency towards theory mapping. Configuration C4 is meant for isolating influence of adding PTM to robust encoder that shows only PTM scores with minimal latency. The configuration C5 assesses if language model and mapping itself is sufficient to match PTM and encoder while it shows

reduction of unwanted API calls using PTM; therefore, there is a minimal effective latency noted. Finally, proposed model with configuration C6 is witnessed to integrate faster encoder, optimal theory mapping, and generative reasoning for practical latency, effective interpretability, and best accuracy.

5. Discussion

In this section, we present a detailed study of the proposed system's outcomes, comparing them to existing state-of-the-art methods on a variety of performance criteria. The talk will focus on the improvements in each metric, as well as their significance in real-world psychological behavior analysis. Figure 4 showcases the outcome of the study graphically for simplifying the interpretation of the outcome in the results discussion. The accomplishment of the result is carried out when both proposed and existing systems are subjected to a uniform assessment environment.

The outcome exhibited in Fig. 4 showcases the potential improvement of accuracy as core performance parameters to identify the psychological state of the subject by proposed system. The outcome showcases the robustness and sustainability of outcome of proposed system in extensive test cases in contrast to the existing system. However, apart from accuracy as performance metric, proposed system has excelled better performance in another performance metric too. The following is the discussion:

- *Interpretation of Accuracy:* The suggested approach regularly outperformed all other datasets in terms of accuracy. For example, using the DAIC-WOZ dataset, the suggested system had an accuracy of 91.5%, which was 6.2% higher than SVM (85.3%) and 4.3% higher than Random Forest (87.2%). Similarly, on the EmoReact dataset, the system achieved 93.1% accuracy, surpassing CNN (85.7%) and RNN (86.4%) by a large margin.
- *Novelty/justification:* The excellent accuracy is due to the combination of Large Language Models (LLMs) such as bidirectional encoder and language model with psychological theory mapping. Unlike traditional models, the proposed framework incorporates behavioral theories (such as cognitive distortions and the Big Five traits) into the decision-making process, allowing it to capture subtle and complex psychological cues that traditional ML models like SVM or Random Forest frequently miss. This eventually attributes to improved generalization over a larger scope of psychological activities during the mechanism of investigating the data from different sources such as social media post and therapy-based transcripts.
- *Quantified outcomes:* The analytical outcome obtained from DAIC-WOZ dataset shows proposed system to score more than 90% for both DAIC-WOZ and EmoReact dataset. It is not be noted that there is no notable bigger difference in outcome of accuracy for any of the dataset that has been tested on proposed scheme. However, for the sake of quantification, it can be stated that there is a range of 4–6% of improvement towards accuracy of proposed system in contrast to existing standard models.
- *Interpretation of Precision:* From the perspective of precision metric, it is seen that proposed system excelled much better performance in contrast to conventional language model or any typical

encoder system.

- *Novelty/justification*: The key reason for better precision score for proposed system can be reasoned with adoption of domain-specific language and fine-tuning while performing training operation. Joint effort of text analysis and psychological theory has led the proposed model to detect various psychological activities without any feasibility to extract any outliers. This is quite important when it concerns to data of mental health as a presence of subtle fragment of outliers can lead to sub-optimal diagnosis.
- *Quantified outcomes*: The proposed scheme has excelled around 89% of precision with CLPsych 2015 dataset while conventional methods have witnessed 80 and 82% in perspective of bidirectional encoder and language model respectively. The score of quantified improvement of proposed study ranges from 5–9% in contrast to standard model exhibiting the reliability associated with detected states of user.
- *Interpretation of Recall*: The trend of outcome of proposed system towards recall as performance metric is nearly similar to that witnessed in study of precision metric
- *Novelty/justification*: Assessment of this performance metric is quite important as ensures the capability of a system to extract highly relevant instances. The enriched outcome in favor of proposed system recall score is due to the adoption of a comprehensive method. This method integrates psychological theory mapping with sophisticated NLP methodologies with an agenda to enhance the detection of subtle indicators of disorder related to mental health. On the contrary, conventional methodologies completely overlook this operation leading to lower recall score.
- *Quantified outcomes*: The proposed model scored 92.1% of recall when subjected to DAIC-WOZ dataset. Random Forest is noted to score 87.5% while SVM is noted to score 85.1%. Therefore, proposed model witnesses around 5–7% of enhanced recall in contrast to existing models.
- *Interpretation of F1-Score*: An extensively enhanced F1-score has been accomplished by the proposed study model exhibiting balanced behaviour towards both recall and precision in contrast to conventional model.
- *Novelty/justification*: From the perspective of model reliability and dependability, F1-score outcomes is very important consideration. This holds valid when subjected to imbalanced classes within any dataset. This is accomplished by adopting psychological theory to finetune and enrich the prediction depending upon the practical and contextual information of disorders related to mental health. Hence, a better balanced performance is noted in proposed model that can be stated to consistently extract all the essential information necessary for reliable prediction.
- *Quantified outcomes*: The proposed model excelled 92.9% of F1-Score when subjected to EmoReact dataset which is the highest even as contrast to highly competent CNN model that has scored only 83.6% of F1-Score. The F1-score for RNN is noted to be 85.1% that is further lower than CNN model. Therefore, the quantified enhancement of F1-Score metric noted by proposed system fall in range of 7–9%.

- *Interpretation of AUC-ROC (Area Under the Curve)*: The AUC-ROC for proposed system is noted to be quite higher for all the dataset. With a minor insignificant variation in the outcome, there is a bigger set of consistency exhibited by proposed model which attributes towards higher applicability in perspective of real-world application.
- *Novelty/justification*: The better score of AUC-ROC can be reasoned with the higher capability of identifying essential feature vectors when subjected to variable / dynamic instances of event. Further, proposed model can perform parallel learning for various psychological states associated with significant behaviour which not only makes the system highly dependable but also faster processing. The inclusion of bidirectional encoder offers more robustness in the proposed model while similar trend is also noted when proposed model is incorporated with language model. When the data is further adjusted based on psychological theory mapping, the outcome of proposed system offers higher classification accuracy. Hence, the system can easily differentiate the frequently adopted or sophisticated emotional indicators that yields to higher score of reliability for application in practical world.
- *Quantified outcomes*: The proposed system has exhibited 5–7% of overall enhanced AUC-ROC in contrast to conventional models. Much better outcome has been noted in DAIC-WOZ dataset where 91.4% of AUC-ROC is noted for random forest while SVM is noted with 89.2%. Proposed system model excelled better results by scoring 94.8% in contrast to existing system.
- *Interpretation of Inference Time*: The suggested system has short inference times (140ms to 170ms) for producing predictions. This is especially applicable in context of clinical settings where the decision-making is carried out by real-time processing. The conventional language model consumes 280ms of inference time which is much longer in contrast to proposed study model. The proposed model is noted to offer a good balance between faster decision and higher performance.
- *Novelty/justification*: It is to be noted that there are certain degree of inclusion of complexity associated with language models and bidirectional encoders and yet their inclusion in proposed system is noted with optimized result. The mechanism adopted in proposed system offers its applicability on time-sensitive configurations such as therapy session, diagnostic, mental health etc. These are certain scenarios that demands faster responsiveness that the proposed model is successful in offering much better in contrast to existing system.
- *Quantified outcomes*: The proposed model is noted to possess 140-170ms of processing time while an individual language model is noted to consume 280ms. As the inference time of proposed system is quite faster in contrast to others, therefore, proposed model is highly appropriate for real-world applications where it bypasses an individual language model by more than 100ms.

The key rationale for the selection of the LLM in the proposed study are as follows:

- *Managing data as natural language*: The analysis of psychological behaviour, especially when it is in the form of a text, it involves different form of an unstructured data. The reason for this unstructured data could be because of inclusion of live post in social media, open-end survey solution, therapy of transcripts, etc. There is an involvement of manual feature extraction process in conventional ML

models which doesn't only consume time but also overlooks complicated differences in the linguistics too. On the contrary, the proposed system adopts language models and bidirectional encoder to facilitate efficient performance for realizing context, semantics, cues of subtle languages, etc in form of natural language. Owing to the inclusion of all these characteristic, the proposed system is highly appropriate for real-time deployment towards investigating psychological traits within behaviour.

- *Enriched Insights with Interpretability*: Another prime reason for opting LLM model is its capability to offer better interpretability. Conventional approaches offer prediction without facilitating rationale for the selection of specific methods. The proposed model uses the standard theory of psychology in the LLM model that not only facilitates behavior classification but also provides enriched context and realization of constructs of the psychological state of an individual. This feature, therefore, assists the model towards a clear understanding of the individual states.
- *Adoption of Standards of Psychology*: The proposed study refers to the usage of the Big Five personality traits and cognitive distortion as the standards of psychological behavior mapping. This potentially contributes towards interpretability without using complexities. Pretrained with a massive and diverse set of data, bidirectional encoder and language model facilitate towards acquiring important patterns in text. The models are further adjusted with domain-specific data (content of social media, transcripts of therapy) as well as integrate standard theory of psychology maps. This potentially assists in classification with better accuracy and interpretability. Hence, the adoption of powerful language models fused with theory-driven context makes the proposed study model different from all existing approaches in the literature. In order to substantiate our discussion of the outcome, we further present a confusion matrix and qualitative error analysis. This will further offer more clearer and better understanding of the extent of the capability of the proposed study.

Confusion matrix has been studied for all the datasets, which are more the nearly similar to each other; therefore, Fig. 5 represents the outcome obtained from the sampled DAIC-WOZ dataset only. The outcome showcases the usual patterns of misclassification. It is been noted that the performance of the model is found to be good when it comes to identifying the states of “*anxiety*” or “*depression*”, it is noted to frequently confuse the state of “*stress*” with that of “*emotional discomfort*”. This happened due to the overlapping of expression of mild symptoms with language patterns. The conclusive statement of this observation from confusion matrix are as follows: i) there is a partial classification overlapping noted for “*stress*” and “*anxiety*”, ii) there is misclassification of “*emotional exhaustion*” with “*mild depression*” sometimes, iii) the sentences with either sarcastic or informal languages (especially from social media) has higher number of false positives.

The outcome of the proposed study has also been assessed for qualitative error cases. Table 8 highlights the cases of qualitative error with certain examples for understanding the misclassified inputs with the prediction of the model.

Table 8
Qualitative error cases

Input Text	True Label	Predicted Label	Explanation
<i>"I feel like I can't breathe when I talk to people"</i>	Anxiety	Stress	Subtle difference misread as stress symptom
<i>"I've just been feeling numb lately. Not happy or sad. Just...nothing."</i>	Depression	Emotional Discomfort	Emotional flatness is hard to distinguish
<i>"I guess I'm fine. Not like anyone cares anyway."</i>	Depression	General	Use of sarcasm confuses the classifier

The information exhibited in Table 8 showcases the need for possible improvement in understanding the context explicitly while determining ambiguous language, sarcasm, or subtle tone. There is a better possibility to mitigate these types of misclassification using some existing methods of detecting sarcasm or performing contextual embedding. The conducted qualitative error analysis showcases the prime areas, such as the interpretation of sarcastic tone and overlapping of emotional categories, that can be taken up for future work direction.

6. Conclusion

The proposed study presents a novel framework that involves framing a psychological behaviour modelling by unifying language model and bidirectional encoder. The presented scheme is meant to contribute towards the area of psychological behavioral analysis by offering an efficient and effective mechanism for determining psychological actions obtained from textual data. This extracted data is then mapped to conventional standard psychological theory. The outcome of the study shows that proposed model has exhibited with better accuracy performance along with lightweight computational efficiency as noted from higher mapping accuracy and faster real-time processing score. The unification of behavior classification and psychological theory mapping results in higher later of an interpretability thereby rendering the study findings highly helpful to the similar researchers as well as therapist. The study has identified various shortcomings e.g. lack of theoretical context and slow processing time while they are addressed in present study to offer enriched model. The outcome of the study also shows that it is highly agile as well as scalable that makes it suitable for being adopted in different types of applications ranging from research work in behavioral science and diagnostic of mental health. The learning outcome of this model is that integration of NLP technique with varied form of psychological theories can significantly enhance interpretability, efficiency, accuracy, etc.

The future direction of the work can be carried out by incorporating multimodal data in current form of system framework itself. The presents state of proposed model is applicable more on textual data; however, in reality, there are inclusion of various data formats e.g., psychological signs, facial expression, voice tone, etc. All these multimodal data will offer more edge to the investigation of indicators of mental health as well as it also involves various complications too. Hence, first stage of future direction

is towards integrating multimodal data where simultaneous processing of various modalities of data is carried out, integrated, and analyzed for more better version of behavioral investigation.

Declarations

Funding: The current research work doesn't involve any form of financial support that could possibly impacted the accomplished outcome.

Data Availability: The dataset used during the current study is available on reasonable request from the first author.

Human/Animal Involvement: We declare that no human or animal involvement is associated with this project.

Clinical trial: Not Applicable.

Code Availability: Not applicable.

Conflict of Interest: We wish to confirm that no known conflicts of interest are associated with this publication.

Ethical Approval: Not applicable.

Consent to Participate: Not applicable.

Consent for Publication: Not applicable.

Author Contributions: Conceptualization, M.K. and A.K.; methodology, A.K.S and .N.G; software, A.K.; validation, A.K, and A.S; formal analysis, M.K.; investigation, A.K.; resources, A.K.S; data curation, A.S.; writing—original draft preparation, M.K.; writing—review and editing, A.S.; visualization, N.G; supervision, G.K.;

References

1. T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, "Natural language processing applied to mental illness detection: a narrative review," *NPJ Digit. Med.*, vol. 5, no. 1, p. 46, 2022, doi.org/10.1038/s41746-022-00589-7.
2. R. Salas-Zárate, G. Alor-Hernández, M. A. Paredes-Valverde, M. del P. Salas-Zárate, M. Bustos-López, and J. L. Sánchez-Cervantes, "Mental-health: An NLP-based system for detecting depression levels through user comments on twitter (X)," *Mathematics*, vol. 12, no. 13, p. 1926, 2024, doi.org/10.3390/math12131926.
3. M. Malgaroli, T. D. Hull, J. M. Zech, and T. Althoff, "Natural language processing for mental health interventions: a systematic review and research framework," *Transl. Psychiatry*, vol. 13, no. 1, p. 309,

2023, doi.org/10.1038/s41398-023-02592-2.

4. A. Montejo-Ráez, M. D. Molina-González, S. M. Jiménez-Zafra, M. Á. García-Cumbreras, and L. J. García-López, "A survey on detecting mental disorders with natural language processing: Literature review, trends and challenges," *Comput. Sci. Rev.*, vol. 53, no. 100654, p. 100654, 2024, doi.org/10.1016/j.cosrev.2024.100654.
5. Z. Hussain, M. Binz, R. Mata, and D. U. Wulff, "A tutorial on open-source large language models for behavioral science," *Behav. Res. Methods*, vol. 56, no. 8, pp. 8214–8237, 2024, doi.org/10.3758/s13428-024-02455-8.
6. A. Sohail and L. Zhang, "Using large language models to facilitate academic work in the psychological sciences," *Curr. Psychol.*, vol. 44, no. 9, pp. 7910–7918, 2025, doi.org/10.1007/s12144-025-07438-2.
7. X. Duan, S. Li, and Z. G. Cai, "MacBehaviour: An R package for behavioural experimentation on large language models," *Behav. Res. Methods*, vol. 57, no. 1, p. 19, 2024, doi.org/10.3758/s13428-024-02524-y.
8. S. Inamdar, R. Chapekar, S. Gite, and B. Pradhan, "Machine learning driven mental stress detection on Reddit posts using Natural Language Processing," *Hum-Cent Intell Syst*, vol. 3, no. 2, pp. 80–91, 2023, doi.org/10.1007/s44230-023-00020-8.
9. P. Kumar, "Large language models (LLMs): survey, technical frameworks, and future challenges," *Artif. Intell. Rev.*, vol. 57, no. 10, 2024. doi.org/10.1007/s10462-024-10888-y
10. T. Shahzad, T. Mazhar, M. U. Tariq, W. Ahmad, K. Ouahada, and H. Hamam, "A comprehensive review of large language models: issues and solutions in learning environments," *Discov. Sustain.*, vol. 6, no. 1, 2025. doi.org/10.1007/s43621-025-00815-8
11. C. Blease and A. Rodman, "Generative artificial intelligence in mental healthcare: An ethical evaluation," *Curr. Treat. Options Psychiatry*, vol. 12, no. 1, 2024. doi.org/10.1007/s40501-024-00340-x
12. E. Gutierrez, W. Karwowski, K. Fiok, M. R. Davahli, T. Liciaga, and T. Ahram, "Analysis of human behavior by mining textual data: Current research topics and analytical techniques," *Symmetry (Basel)*, vol. 13, no. 7, p. 1276, 2021, https://doi.org/10.3390/sym13071276
13. H. Uchida, "Equivalence of paper and smart-phone versions of the Beck Depression Inventory-II," *J. Clin. Med*, vol. 14, no. 2, 2025. doi.org/10.3390/jcm14020500
14. H. Chung et al., "Diagnosis of pain deception using Minnesota Multiphasic Personality Inventory-2 based on XGBoost machine learning algorithm: A single-blinded randomized controlled trial," *Medicina (Kaunas)*, vol. 60, no. 12, 2024. doi.org/10.3390/medicina60121989
15. M. Jain et al., "Speech emotion recognition using Support Vector Machine," *arXiv [eess.AS]*, 2020. http://arxiv.org/abs/2002.07590
16. M. Z. Uddin, K. K. Dysthe, A. Følstad, and P. B. Brandtzaeg, "Deep learning for prediction of depressive symptoms in a large textual dataset," *Neural Comput. Appl.*, vol. 34, no. 1, pp. 721–744, 2022. doi.org/10.1007/s00521-021-06426-4

17. M. Kerasiotis, L. Ilias, and D. Askounis, "Depression detection in social media posts using transformer-based models and auxiliary features," *Soc. Netw. Anal. Min.*, vol. 14, no. 1, 2024, doi.org/10.1007/s13278-024-01360-4.
18. M. R. R. Rana, A. Nawaz, S. U. Rehman, M. A. Abid, M. Garayevi, and J. Kajanová, "BERT-BiGRU-senti-GCN: An advanced NLP framework for analyzing customer sentiments in E-commerce," *Int. J. Comput. Intell. Syst.*, vol. 18, no. 1, 2025, doi.org/10.1007/s44196-025-00747-1.
19. K. S. L. Prasanna, S. Lokesh, G. Chandramouli, N. Vimala, M. Pu-Neeth, and P. D. S. S. Chandranath, "BERT-QA: Empowering intelligent question answering with NLP and entity recognition," in *3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, 2024, pp. 1006–1010, 10.1109/ICAAIC60222.2024.10575242.
20. M. M. Abdelgwad, T. H. A. Soliman, and A. I. Taloba, "Arabic aspect sentiment polarity classification using BERT," *J. Big Data*, vol. 9, no. 1, 2022, doi.org/10.1186/s40537-022-00656-6.
21. K.-S. Park and H. Choi, "How to harness the power of GPT for scientific research: A comprehensive review of methodologies, applications, and ethical considerations," *Nucl. Med. Mol. Imaging (2010)*, vol. 58, no. 6, pp. 323–331, 2024, doi.org/10.1007/s13139-024-00876-z.
22. R. S.-Y. Wong, "ChatGPT in psychiatry: promises and pitfalls," *Egypt. J. Neurol. Psychiatr. Neurosurg.*, vol. 60, no. 1, 2024, doi.org/10.1186/s41983-024-00791-2.
23. L. Bojić, I. Stojković, and Z. Jolić Marjanović, "Signs of consciousness in AI: Can GPT-3 tell how smart it really is?," *Humanit. Soc. Sci. Commun.*, vol. 11, no. 1, 2024, doi.org/10.1057/s41599-024-04154-3.
24. M. T. Rietberg, V. B. Nguyen, J. Geerdink, O. Vijlbrief, and C. Seifert, "Accurate and reliable classification of unstructured reports on their diagnostic goal using BERT models," *Diagnostics (Basel)*, vol. 13, no. 7, p. 1251, 2023, doi.org/10.3390/diagnostics13071251.
25. F. Neha, D. Bhati, D. K. Shukla, and M. Amiruzzaman, "ChatGPT: Transforming healthcare with AI," *AI (Basel)*, vol. 5, no. 4, pp. 2618–2650, 2024, doi.org/10.3390/ai5040126.
26. T. Lai et al., "Supporting the demand on mental health services with AI-based conversational large language models (LLMs)," *BioMedInformatics*, vol. 4, no. 1, pp. 8–33, 2023, doi.org/10.3390/biomedinformatics4010002.
27. I. Levkovich, "Evaluating diagnostic accuracy and treatment efficacy in mental health: A comparative analysis of large language model tools and mental health professionals," *Eur. J. Investig. Health Psychol. Educ.*, vol. 15, no. 1, 2025, doi.org/10.3390/ejihpe15010009.
28. E. M. Abdelaziz et al., "Effectiveness of cognitive behavioral therapy (CBT) on psychological distress among mothers of children with autism spectrum disorder: The role of problem-solving appraisal," *Behav. Sci. (Basel)*, vol. 14, no. 1, 2024, doi.org/10.3390/bs14010046.
29. M.-N. Yu, Y.-N. Chang, and R.-H. Li, "Relationships between Big Five personality traits and psychological well-being: A mediation analysis of social support for university students," *Educ. Sci. (Basel)*, vol. 14, no. 10, p. 1050, 2024, doi.org/10.3390/educsci14101050.

30. Z. Han, S. Chen, Y. Zhou, Y. Liu, and C. Guo, "The longitudinal relationship between self-esteem, life satisfaction, and depressive and anxiety symptoms among Chinese adolescents: Within- and between-person effects," *Behav. Sci. (Basel)*, vol. 15, no. 2, p. 182, 2025, doi.org/10.3390/bs15020182.
31. A. Hutto et al., "Using a natural language processing toolkit to classify electronic health records by psychiatric diagnosis," *Health Informatics Journal*, no. 4, 2024, doi.org/10.1177/14604582241296411.
32. M. M. Hossain, M. S. Hossain, M. F. Mridha, M. Safran, and S. Alfarhood, "Multi task opinion enhanced hybrid BERT model for mental health analysis," *Sci. Rep.*, vol. 15, no. 1, p. 3332, 2025, doi:10.1038/s41598-025-86124-6.
33. H. Na, "CBT-LLM: A Chinese large language model for Cognitive Behavioral Therapy-based mental health question answering," *arXiv [cs.CL]*, 2024, doi.org/10.48550/arXiv.2403.16008.
34. M. Zolfaghari, Y. Zhu, P. Gehler, and T. Brox, "CrossCLR: Cross-modal contrastive learning for multi-modal video representations," pp. 1450–1459, *arXiv [cs.CV]*, 2021.
35. J. Li et al., "Assessing severity of pediatric pneumonia using multimodal transformers with multi-task learning," *Digital Health*, vol. 10, 2024, doi:10.1177/20552076241305168.
36. "DAIC-WOZ database," *DAIC-WOZ*, 16-Feb-2022. [Online]. Available: <https://dcapswoz.ict.usc.edu>. [Accessed: 29-Aug-2025].
37. İ. Baydili, B. Tasci, and G. Tasci, "Deep learning-based detection of depression and suicidal tendencies in social media data with feature selection," *Behavioral Sciences*, vol. 15, no. 3, 2025, doi:10.3390/bs15030352.
38. P. Buddhitha and D. Inkpen, "Multi-task learning to detect suicide ideation and mental disorders among social media users," *Frontiers in Research Metrics and Analytics*, vol. 8, 2023, doi:10.3389/frma.2023.1152535.
39. S. MacAvaney, A. Mittu, G. Coppersmith, J. Leintz, and P. Resnik, "Community-level research on suicidality prediction in a secure environment: Overview of the CLPsych 2021 shared task," in *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, 2021, doi:10.18653/V1/2021.CLPSYCH-1.7.
40. S. Ji, X. Li, Z. Huang, and E. Cambria, "Suicidal ideation and mental disorder detection with attentive relation networks," <http://arxiv.org/abs/2004.07601>, 2020.
41. A. Kumar, M. Kumar, V. P. Bhardwaj, S. Kumar, and S. Selvarajan, "A novel skin cancer detection model using modified finch deep CNN classifier model," *Sci. Rep.*, vol. 14, no. 1, p. 11235, 2024, doi.org/10.1038/s41598-024-60954-2.
42. A. Kumar et al., "Flamingo-optimization-based deep convolutional neural network for IoT-based arrhythmia classification," *Sensors (Basel)*, vol. 23, no. 9, p. 4353, 2023, doi.org/10.3390/s23094353.
43. M. Kumar et al., "Healthcare Internet of Things (H-IoT): Current trends, future prospects, applications, challenges, and security issues," *Electronics (Basel)*, vol. 12, no. 9, p. 2050, 2023, doi.org/10.3390/electronics12092050.

44. I. de la Cuétara, "Emotional development in people with high capacities: Induction of emotions through pictorial abstraction," *Sustainability*, vol. 12, no. 15, p. 5912, 2020, doi.org/10.3390/su12155912.
45. L. Ningjian, "Cannon-bard theory of emotion," in *The ECPH Encyclopedia of Psychology*, Singapore: Springer Nature, 2024, pp. 1–2.
46. G. Šimić et al., "Understanding emotions: Origins and roles of the amygdala," *Biomolecules*, vol. 11, no. 6, p. 823, 2021, doi.org/10.3390/biom11060823.
47. A. P. Natoli, "The DSM's reconnection to psychoanalytic theory through the Alternative Model for Personality Disorders," *J. Am. Psychoanal. Assoc.*, vol. 67, no. 6, pp. 1023–1045, 2019, doi.org/10.1177/0003065120903060.
48. "CLPsych 2015 shared task evaluation," *Jhu.edu*. [Online]. Available: <https://www.cs.jhu.edu/~mdredze/clpsych-2015-shared-task-evaluation/>. [Accessed: 29-Aug-2025].
49. B. Nojavanasghari, T. Baltrušaitis, C. E. Hughes, and L.-P. Morency, "EmoReact: a multimodal approach and dataset for recognizing emotional responses in children," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, p. 137–144, 2016, doi.org/10.1145/2993148.2993168. <http://multicomp.cs.cmu.edu/resources/emoreact-dataset/>
50. Mumtaz, Shahid, et al. "Guest Editorial: Introduction to the Special Section on Large Models for Future Vehicles and Transportation." *IEEE Transactions on Vehicular Technology* 74.2 (2025): 1827–1829.
51. Salih, Abdulrahman, Xiaoqi Ma, and Evtim Peytchev. "Implementation of hybrid artificial intelligence technique to detect covert channels attack in new generation internet protocol IPv6." *Leadership, Innovation and Entrepreneurship as Driving Forces of the Global Economy: Proceedings of the 2016 International Conference on Leadership, Innovation and Entrepreneurship (ICLIE)*. Cham: Springer International Publishing, 2016.
52. Xu, Guoji, Jianhua Chen, and Qin Chen. "Application of artificial neural networks to wave load prediction for coastal bridges." *Proceedings of the 3rd International Conference on Communication and Information Processing*. 2017.
53. Broni-Bedaiko, Clifford, et al. "El Niño-Southern Oscillation forecasting using complex networks analysis of LSTM neural networks." *Artificial Life and Robotics* 24.4 (2019): 445–451.
54. Wu, Zonghan, et al. "A comprehensive survey on graph neural networks." *IEEE transactions on neural networks and learning systems* 32.1 (2020): 4–24.
55. Rutinowski, Jérôme, et al. "The self-perception and political biases of ChatGPT, arXiv." *arXiv preprint arXiv:2304.07333* (2023).
56. Al-Saggaf, Yeslam, and Rachel Hogg. "The effect of children's phubbing on parents' psychological wellbeing: a moderated mediation analysis." *Human Behavior and Emerging Technologies* 2024.1 (2024): 9719351.
57. Sangka, Khresna Bayu, Ramadhan Defitri Pratama, and Rahmad Wahyu Hidayat. "Redefining Education in the Digital Age: The Role of GPT Models in Academic Fraud and Logistics." *Human*

Figures

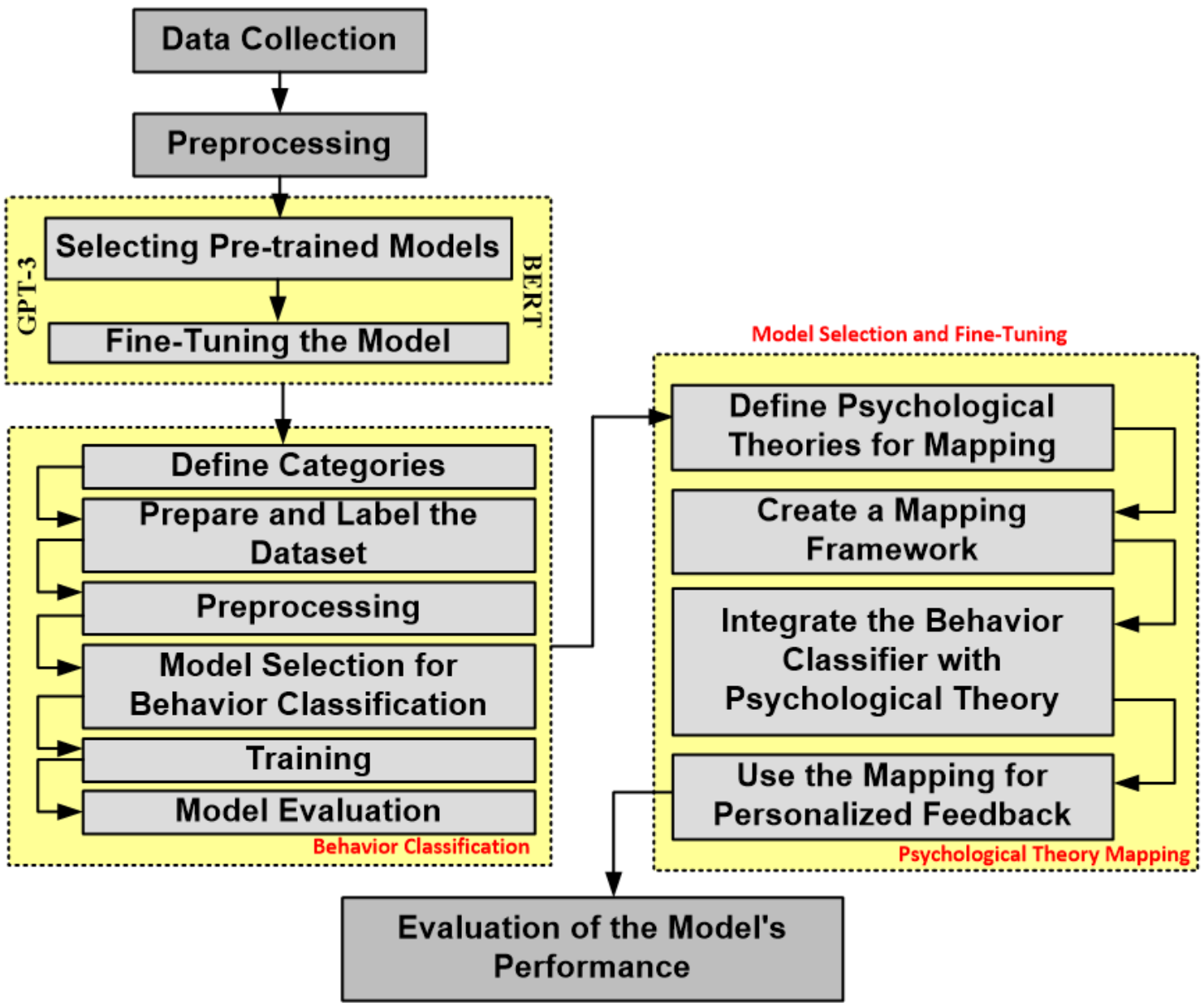


Figure 1
Proposed architecture

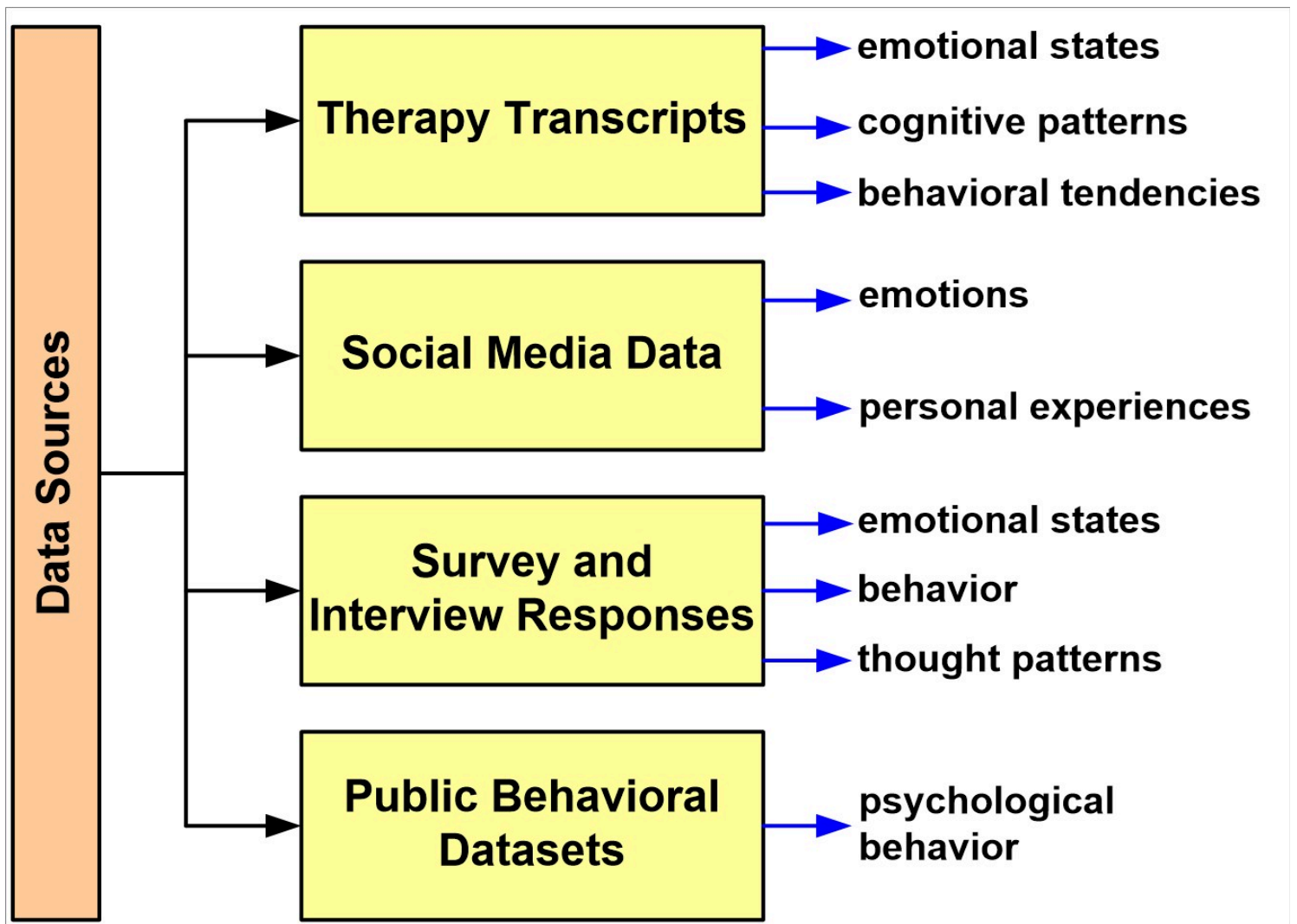


Figure 2

Data Sources involved in the study

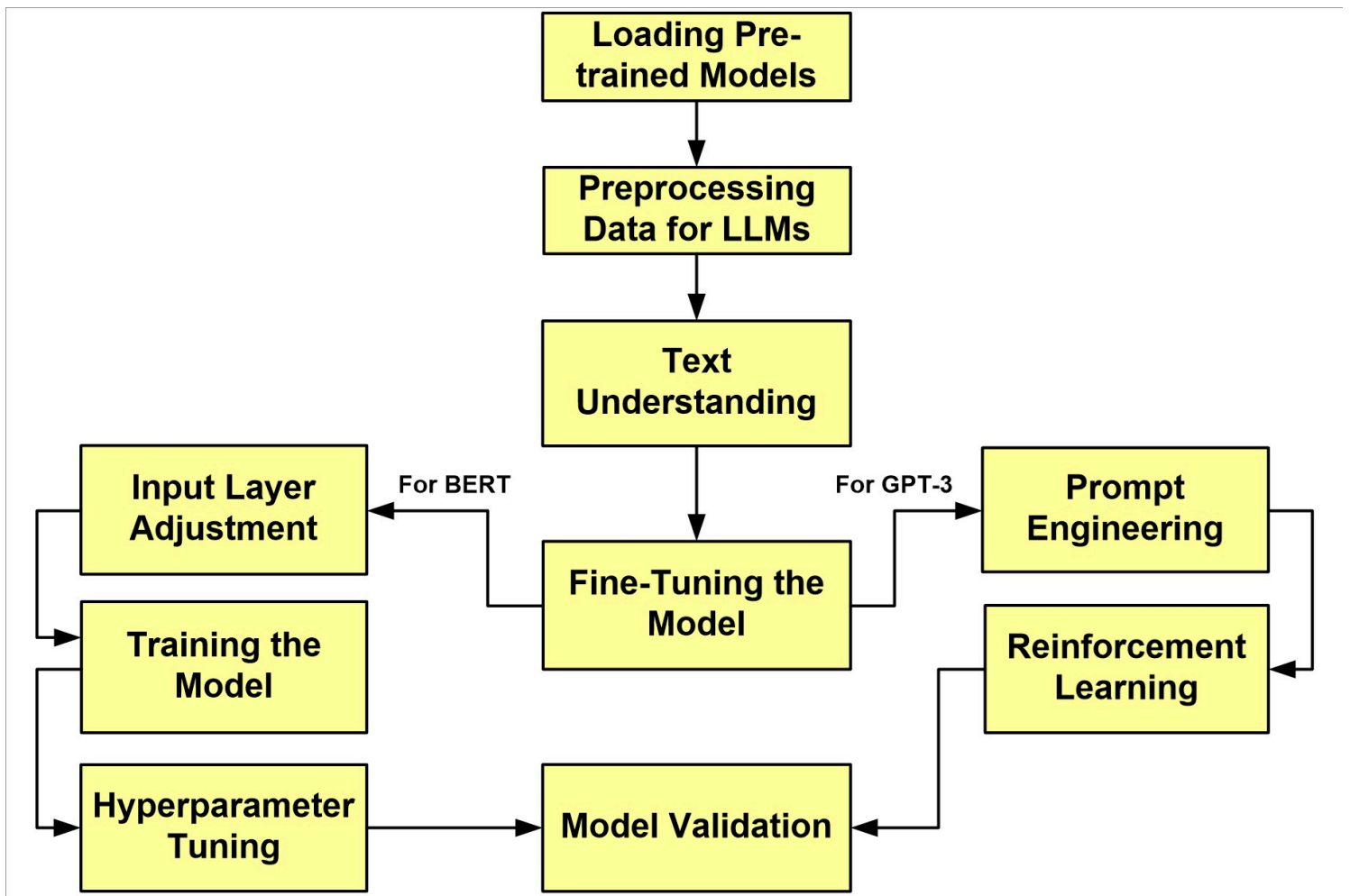
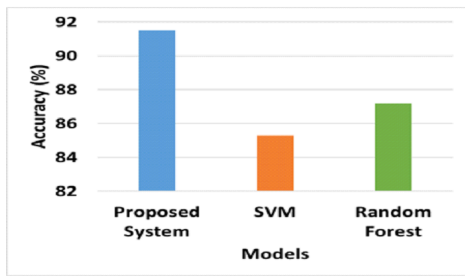
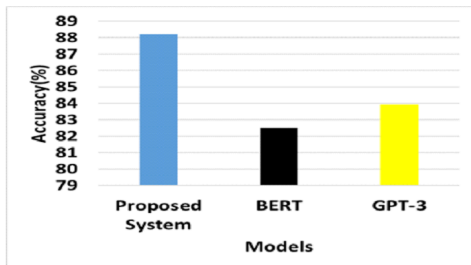


Figure 3

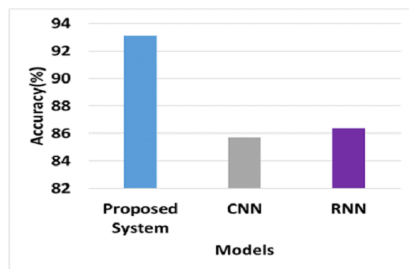
Mechanism involved in pre-trained model selection



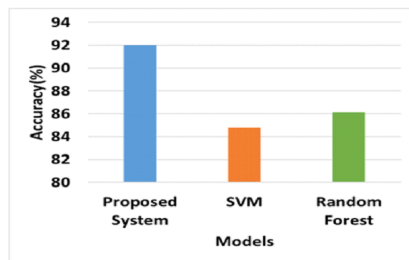
(a)



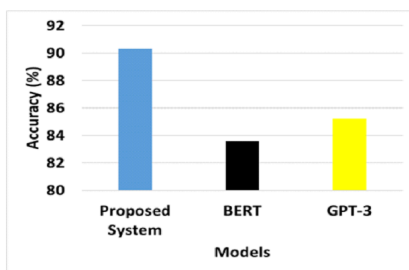
(b)



(c)



(d)



(e)

Figure 4

Graphical outcome of study

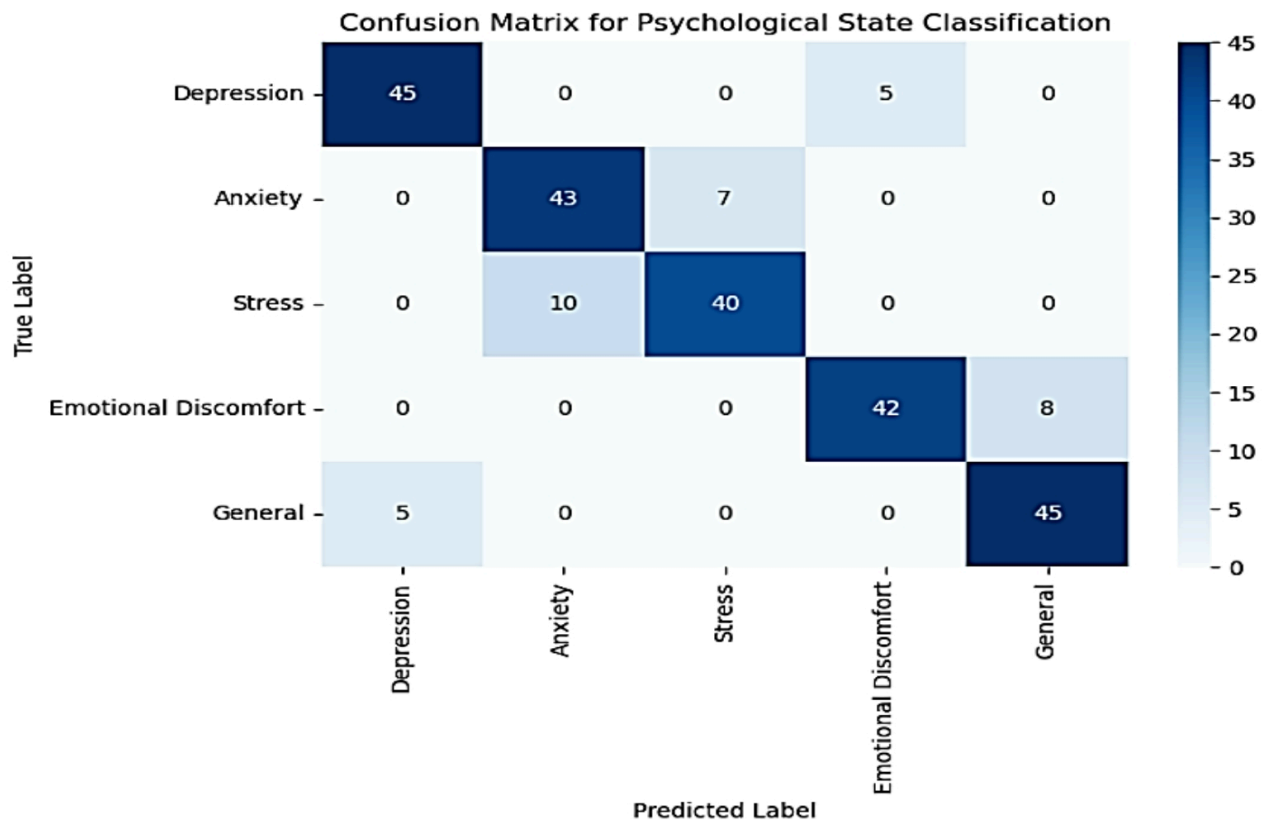


Figure 5

Confusion Matrix for classifying psychological state