Transforming Mental Health Care: Harnessing the Power of RoBERTa for Assessing and Supporting Anxiety, Stress, and Depression

Aarav Sureban

Received July 13, 2023
Accepted November 06, 2023
Electronic access November 15, 2023

This study aims to develop a classification model using RoBERTa, a pre-trained language model to predict levels of depression, anxiety, and stress. The dataset comprises 39,776 responses collected with an online version of the Depression Anxiety Stress Scales (DASS) between 2017 and 2019. Precision, Recall, and F1 scores were used to evaluate the model's performance, indicating its efficacy in accurately identifying mental health conditions. The RoBERTa model outperformed traditional machine learning algorithms, showcasing its advanced language modeling capabilities and ability to capture context-specific information. Its Precision, Recall, and F1 score for depression classification ranged from 0.925 to 0.959, 0.926 to 0.962, and 0.926 to 0.960, respectively. For anxiety classification, the model achieved Precision values between 0.937 and 0.945, Recall values between 0.908 and 0.946, and F1 scores between 0.937 and 0.944. In stress classification, the Precision ranged from 0.912 to 0.929, Recall from 0.908 to 0.924, and F1 score from 0.910 to 0.923. These results highlight the RoBERTa model's exceptional performance in accurately classifying instances of different severity levels across depression, anxiety, and stress. Future research should focus on cross-validation, external dataset evaluation, and parameter tuning to enhance generalizability.

Introduction

Stress, anxiety, and depression are prevalent mental health issues significantly affecting individuals' well-being. Stress is a natural response to challenging or demanding situations that can be overwhelming. It arises from various sources and factors, such as academic responsibilities, work-related pressures, financial difficulties, or interpersonal conflicts. While short-term stress can be manageable, long-term stress can have detrimental effects on both mental and physical health and should not be ignored. Anxiety can be described as persistent feelings of fear, worry, or unease. It can develop from a temporary emotional response to a disorder that interferes with everyday life. Anxiety disorders exhibit a range of conditions, such as generalized anxiety disorder (GAD), panic disorder, social anxiety disorder, and various phobias. These disorders have many symptoms, such as shortness of breath, trembling, and excessive sweating. Finally, depression is a mood disorder encompassing persistent sadness, loss of interest in activities, and several cognitive/physical symptoms. As seen in Figure I, almost 35% of people around the world suffer from depression or anxiety, with almost 50% of 18-29-year-olds suffering from these dangerous disorders, showing its severity. Symptoms include a lack of energy, changes in appetite and sleep patterns, difficulty concentrating, and even thoughts of suicide, all of which substantially impact daily function and life.

However, one significant challenge individuals can face with these issues is that they find it difficult to openly discuss their struggles with others[1]. The stigma surrounding mental health and the fear of being judged or made fun of can create barriers to seeking help and support, with individuals even worrying about burdening others with their problems. On top of this, expressing their emotions and conditions to others is highly subjective and difficult to convey accurately. As a result, individuals tend to suffer from their problems in silence, feeling isolated and disconnected from others. However, sentiment analysis can be valuable and essential in understanding these struggles. By analyzing the sentiments expressed in conversations about mental health, researchers and medical professionals can gain insight into these emotions, such as sadness, loneliness, shame, and vulnerability. This information can be used to develop more focused interventions and support systems to address the issues that individuals face.

In light of these challenges, this research introduces a novel approach to addressing mental health concerns. We aim to leverage the power of state-of-the-art natural language processing models, specifically a pre-trained version of the Robustly Optimized BERT Approach (RoBERTa), in combination with the Depression, Anxiety, and Stress Scale-42 test (DASS-42 test). This approach uniquely combines advanced text analysis with self-report data to detect levels of stress, anxiety, and depression. Unlike traditional methods that may require external equipment or rely solely on linguistic analysis, our approach utilizes the inherent vagueness of the DASS-42 test questions, enabling the network to analyze user responses without collecting personal information directly. Furthermore, we harness an open-source dataset encompassing diverse age groups and geographical re-

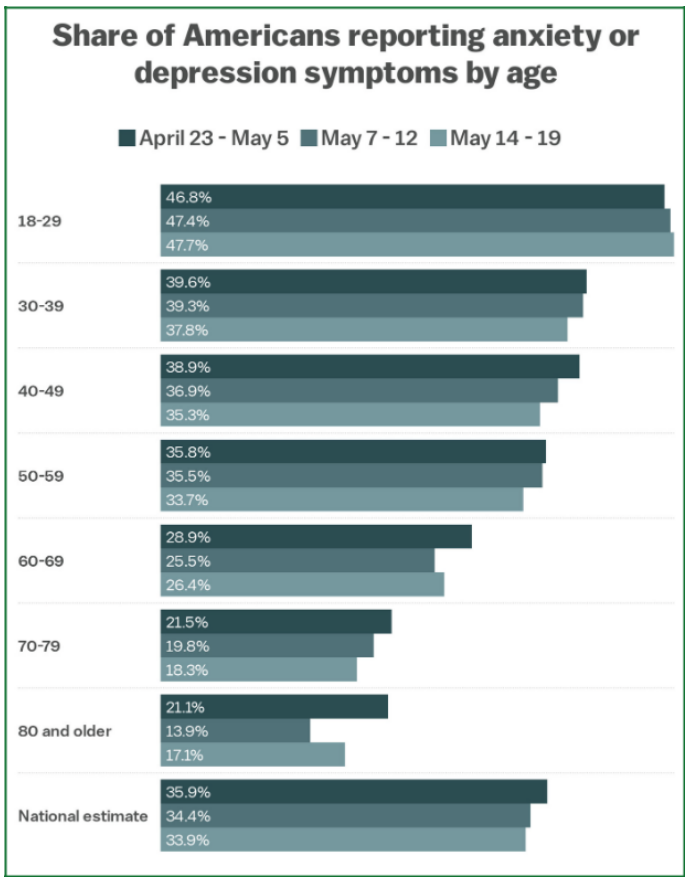

Fig. 1 Percentages of individuals suffering from depression or anxiety by age in the U.S.

gions, significantly increasing the sample size and making our research more broadly applicable.

Past research has been done that focuses on sentiment analysis related to stress, anxiety, and depression, with many contributions and limitations. In[2], Stamatis investigates how the language features of outgoing text messages can predict the mental health conditions of the senders. The authors use several linguistic dictionaries and tools, such as Linguistic Inquiry and Word Count (LIWC), National Research Council (NRC) Emotion Lexicon, and depression and stress dictionaries, to extract language features from text messages. The paper then applies hierarchical linear models (HLM) to evaluate the relationships between language features and symptoms of depression, generalized anxiety, and social anxiety. These multilevel models are used to analyze data with a hierarchical structure, such as when individuals are nested within groups or organizations. HLM allows us to examine relationships and make predictions at the individual and group levels. It considers that individuals within the same group may be more similar to each other than individuals from different groups. By using HLM, researchers in the mentioned paper were able to study how language features in text messages related to symptoms of depression, anxiety, and social anxiety while considering both the unique characteristics of individuals and the collective.

characteristics of the groups they belong to. This paper lacks reliance on self-reported measures of mental health symptoms. The authors used text-message-based sentiment analysis to predict symptoms of depression, generalized anxiety, and social anxiety in a sample of 335 adults over 16 weeks. However, they did not collect any self-report data from the participants to validate their findings or to account for individual differences in language use and expression of emotions. Self-reported measures are widely used in clinical research and practice to assess mental health outcomes and to monitor treatment progress. Therefore, it is unclear how well the text message-based sentiment analysis reflects the actual experiences and perceptions of the participants and how it compares to other established methods of measuring mental health symptoms. On top of this, the paper has limitations due to the lack of a robust data set due to the use of a homogeneous sample of college students and the many factors that may affect text messaging behavior.

In[3], Jalukar and Ratnaparkhi propose a system that can detect

depression from text, video, and audio inputs using sentiment analysis and natural language processing methods. The paper uses various techniques to extract features from different modalities, such as Term Frequency-Inverse Document Frequency (TF-IDF) for text, facial expression recognition for video, and speech emotion recognition for audio, and different classifiers, such as Support Vector Machines (SVM), k-nearest neighbors algorithms (KNN), Naive Bayes, and Random Forests, to classify the inputs as positive or negative. Each input, whether audio or text, was assigned a sentiment score between 0 and 1. 0 meant that the input conveyed negative opinions and emotions, and 1 indicated that the input conveyed positive opinions and emotions. The paper claims that the system can help identify people suffering from depression and provide them with timely intervention and support. However, there is a lack of large-scale and diverse datasets for depression detection.

One of the sources of data that can be used for the detection of stress, anxiety, and depression is physiological signals, such as electrocardiogram (ECG), electroencephalogram (EEG), galvanic skin response (GSR), and heart rate variability (HRV). These signals reflect nervous system activity and can capture the emotional states of an individual. For example, ECG can measure the electrical activity of the heart and indicate changes in heart rate and rhythm due to stress or anxiety. EEG can measure the electrical activity of the brain and indicate changes in brain waves due to mood or attention. GSR can measure the skin's electrical conductance of the skin and indicate changes in sweating due to arousal or emotion. HRV can measure the variation in time intervals between heartbeats and indicate changes in cardiac regulation due to stress or depression.

In[4], Hole and Anand propose a system that can detect human mental stress and emotion using EEG signals. The paper uses a brain-computer interface device to collect EEG signals from participants exposed to different stimuli that induce positive or negative emotions. The signals are classified into four basic emotions: happy, sad, angry, and neutral. However, the complexity and cost of EEG devices make the system inaccessible to most individuals.

While powerful, sentiment analysis poses many challenges, such as sarcasm, irony, and figurative language. Words with multiple definitions can make it difficult for sentiment analysis algorithms to detect their true meaning, providing inaccurate results. On top of this, various sentences and pieces of text can exhibit multipolarity: having multiple sentiments within the same piece of text being analyzed. Sentiment analysis algorithms would have to be able to assign multiple sentiments for one piece of text in order to be effective. Researchers have proposed various methods and techniques to address these challenges, ranging from rule-based and lexicon-based approaches to machine learning and deep learning models.

One of the most popular approaches for sentiment analysis is neural networks. Neural networks are inspired by the structure and function of the human brain, which also consists of billions of neurons that communicate with each other. However, neural networks are not exact replicas of the brain, and they use different algorithms and architectures to achieve their goals. Neural networks can be divided into different types based on their structure and learning method. Some common types are feedforward neural networks (FNNs), where the connections between nodes do not form a cycle, recurrent neural networks (RNNs), where the connections between nodes create loops, allowing for information to be stored, and convolutional neural networks (CNNs), which are feedforward neural networks used on images. Each type has advantages and disadvantages and can be applied to different domains and problems. Neural networks can learn from data and perform various tasks. They comprise interconnected units called neurons that can process information and transmit signals to other neurons. Neural networks can be trained to recognize patterns, classify objects, generate text, play games, and in this case, classify sentiments in stress, anxiety, and depression.

In[5], the authors study various deep neural networks (DNNs) used for sentiment classification and their applications. The paper examines several contemporary DNN models and their theories, such as FNNs, CNNs, RNNs, long short-term memory (LSTMs), which are a type of RNNs that can process sequential data and capture long-term dependencies, gated recurrent units (GRUs), and attention-based models.

In[6], Li and Liu used physiological signals to detect levels of stress, anxiety, and depression. They developed two deep neural networks: a 1-dimensional (1D) CNN and a multilayer perceptron (MLP) neural network. MLPs are neural networks that use fully connected layers to learn global features from input data. They used CNNs to analyze data from chest-worn sensors and MLPs to analyze data from wrist-worn sensors. The sensors collected ECG, GSR, accelerometer, and temperature data from participants who were exposed to different emotional stimuli: baseline (neutral), stress (participants were given a math test), and amusement (participants were given a funny video to watch). The neural networks performed two tasks: binary classification for stress detection (stressed vs. non-stressed) and 3-class classification for emotion classification (baseline vs. stressed vs. amused). The neural networks achieved high accuracy rates for both tasks: 99.80% and 99.55% for CNNs, 99.65%, and 98.38% for MLPs.

In[7], the authors also used physiological signals to detect stress, anxiety, and depression. They proposed an efficient approach using an LSTM-based RNN. The authors used LSTMs to analyze text data from a public online information channel for young people in Norway. The text data consisted of the youth's own questions on various topics, such as health, sexuality, relationships, etc. The authors extracted features from the text data based on the reflection of possible symptoms of depression pre-defined by medical and psychological experts. These fea-

tures were then fed into the LSTM network to classify the text data into depression posts (describing self-perceived symptoms of depression) and non-depression posts (not describing such symptoms). The LSTM network achieved a 94% accuracy rate for depression detection.

A different source of data that can be used to detect stress, anxiety, and depression is speech. Speech can convey information about a person's emotional state through various acoustic features, such as pitch, intensity, duration, and spectral properties. For example, speech can become slower, quieter, lower-pitched, or more monotonous due to depression. Speech can also become faster, louder, higher-pitched, or more variable due to stress or anxiety. One of the studies that used speech for stress, anxiety, and depression detection was conducted by Shenoy and Ghosh[8]. They used CNNs to learn useful characteristics of depression from speech. They used a spectrogram as the input representation of speech. A spectrogram is a visual representation of the frequency spectrum of sound over time. The CNN model was designed to extract features from the spectrogram and classify it into two classes: depressed and non-depressed. The authors used a publicly available dataset called DAIC-WOZ that contained audio recordings of clinical interviews with participants who were diagnosed with depression or not. The CNN model achieved an 83% accuracy rate for depression detection.

In[9], the authors employed an RNN to represent features of video-based input into a deep-learning neural network for better high-quality depression detection. The authors used a publicly available dataset called AVEC 2014 that contained video recordings of participants who were asked to watch emotional clips and answer questions about their moods and feelings. The RNN model achieved an 86% accuracy rate for depression detection.

Video is a third data source that can be used for SAD detection. Video can capture information about a person's facial expressions, gestures, and eye movements that indicate their emotional state. For example, facial expressions can show sadness, anger, fear, happiness, etc., due to different emotions. Due to different attitudes, gestures can show agitation, nervousness, or confidence. Due to different cognitive processes, eye movements can show attention, interest, or boredom. Zhang conducted one of the studies that used video for SAD detection[10]. They proposed a novel framework that combined CNNs and graph convolutional networks (GCNs) to analyze facial expressions for stress detection. CNNs were used to extract local features from facial regions, such as eyes, nose, mouth, etc., while GCNs were used to capture global features from facial landmarks, such as eyebrows, cheeks, and chin, by modeling them as nodes in a graph. The authors used a publicly available dataset called MAHNOB-HCI that contained video recordings of participants exposed to different stress-inducing tasks, such as mental arithmetic or public speaking. The CNN-GCN framework achieved a 91% accuracy rate for stress detection.

In[11], Wang developed a multimodal approach that integrated facial expressions, speech, and text for depression detection. They used a multimodal recurrent neural network (MRNN) to fuse the features from different modalities and learn their temporal dynamics. They also used an attention mechanism to focus on the most relevant features for each modality. The authors used a publicly available dataset called Distress Analysis Interview Corpus (DAIC) that contained video recordings of clinical interviews with participants who were diagnosed with depression or not. The MRNN-attention model achieved an 85% accuracy rate for depression detection.

In light of these current technologies, this research introduces a novel approach to addressing mental health concerns. We aim to leverage the power of state-of-the-art natural language processing models, specifically a pre-trained version of the Robustly Optimized BERT Approach (RoBERTa), in combination with the Depression, Anxiety, and Stress Scale-42 test (DASS-42 test). This approach uniquely combines advanced text analysis with self-report data to detect levels of stress, anxiety, and depression. Unlike traditional methods that may require external equipment or rely solely on linguistic analysis, our approach utilizes the inherent vagueness of the DASS-42 test questions, enabling the network to analyze user responses without collecting personal information directly. Furthermore, we harness an open-source dataset encompassing diverse age groups and geographical regions, significantly increasing the sample size and making our research more broadly applicable.

A transformer model will be used to perform sentiment analysis to address the issues highlighted in these papers. More specifically, a pre-trained version of Robustly Optimized BERT Approach (RoBERTa)[12] will be used and combined with the questions asked in the Depression, Anxiety, and Stress Scale-42 test (DASS-42 test)[13], which is a self-report test that assesses the emotional state of individuals to detect levels of stress, anxiety, and depression. It asks 42 questions, and its results will be compiled and analyzed using a transformer model. It does not require any external equipment, and the vague nature of the questions asked will allow the network to not need or collect personal information from users. On top of this, an open-source data set will be used that has data from all ages and countries, significantly increasing the sample size. The notebook "Terrified— Prediction + Feature Selection"[14] available on Kaggle that we use as a reference, uses this data set to evaluate the accuracy of many models, such as Random Forest, AdaBoost, Gradient Boosting, KNN, Support Vector Classifier (SVC), and Voting Classifier. The transformer model will determine whether RoBERTa's ability to capture relationships within the text can achieve more accurate results than the other models. This research is structured with the following sections: Results, discussion, methods, and conclusion. The results section will present the findings with the transformer model. The discussion section will analyze the results, interpret their implications, and discuss future research opportunities. The methods section will present

Research Paper	Methods and Models Used	Accuracy Rate
Stamatis (In2)	Linguistic dictionaries, HLM	Not reported
Jalukar and Ratnaparkhi (In3)	TF-IDF, Facial Expression Recognition, Speech Emotion Recognition,SVM, KNN, Naive Bayes, Random Forests	Not reported
Hole and Anand (In4)	Brain-computer interface, EEG signals,Emotion classification	Not reported
Li and Liu (In6)	1D CNN, MLP, ECG, GSR, Accelerometer,Temperature data	99.80% (CNN), 99.65% (MLP)
Uddin (In7)	LSTM-based RNN, Text data, Depression post classification	94%
Shenoy and Ghosh (In8)	CNN,Spectrogram,Audio recording,Depression classification	83%
M. Al Jazaery and G. Guo	RNN, Video recordings, Depression detection	86%
Zhang (In10)	CNNs, GCNs, Facial Expressions, Video recordings,Stress detection	91%
Wang (In11)	MRNN,Attention Mechanism,Facial Expressions,Speech, Text,Depression detection	85%

Table 1 State-of-the-art techniques used in past research along with their accuracy rates if reported.

the methods and techniques used, including pre-training and the architecture of the transformer-based sentiment analysis model. Finally, the conclusion will summarize the research and analyze its importance.

Results

The objective of this study was to develop a classification model to predict the levels of depression, anxiety, and stress based on a dataset consisting of 39,776 responses. The model demonstrated high accuracy for all three scales, indicating its efficacy in accurately identifying mental health conditions. The dataset used in this study included various columns representing different aspects of mental health assessment. The "Depression Level," "Anxiety Level," and "Stress Level" columns contained numerical values ranging from 0 to 4, representing the severity of depression, anxiety, and stress symptoms, respectively. The severity levels were categorized as follows:

0: No symptoms
1: Minimal symptoms
2: Mild symptoms
3: Moderate symptoms
4: Severe symptoms

Upon evaluating the model's performance, several metrics were used to assess its effectiveness in classifying the levels of depression, anxiety, and stress. These metrics provide insights into the model's Precision, Recall, and F1 score.

Precision measures the proportion of correctly classified instances among those predicted as a specific severity level. It indicates the model's ability to identify instances of a particular severity level accurately. It is calculated as follows: $(1) Precision = TP/(TP+FP)$

where TP (True Positives) refers to the number of instances that are correctly classified as positive (in this case, instances with a specific severity level), and FP (False Positives) refers to the number of instances that are incorrectly classified as positive when they should have belonged in a different category. The Precision values for depression classification range from 0.925 to 0.959 across different severity levels, with the highest Precision observed for severity level 3 (0.959). For anxiety classification, the Precision values range from 0.937 to 0.945, with severity level 1 achieving the highest Precision (0.945). In stress classification, the Precision values range from 0.912 to 0.929, with severity level 1 achieving the highest Precision (0.929). These Precision values indicate the model's high accuracy in correctly classifying instances within each severity level for depression, anxiety, and stress.

Recall measures the proportion of correctly classified instances among the true instances of a specific severity level. It is calculated as follows: $(2) Recall = TP/(TP+FP)$

where FN (False Negatives) refers to the number of instances that are incorrectly classified as negative (not having the specific severity level) when they should have been classified as positive. It reflects the model's ability to capture and identify all instances of a particular severity level. The Recall values for depression classification range from 0.926 to 0.962 across different severity levels, with the highest Recall observed for severity level 3 (0.962). For anxiety classification, the Recall values range from 0.908 to 0.946, with severity level 3 achieving the highest Recall (0.946). In stress classification, the Recall values range from 0.908 to 0.924, with severity level 3 achieving the highest Recall (0.924). These Recall values indicate that the model effectively captures and identifies instances within each severity level for depression, anxiety, and stress.

The F1 score is a metric that combines Precision and Recall into a single value, providing an overall assessment of the model's performance. It is calculated as follows:

$(3) F1score = 2 * (Precision * Recall)/(Precision + Recall)$

The F1 scores for depression classification range from 0.926 to 0.960 across different severity levels, with the highest F1 score observed for severity level 3 (0.960). For anxiety classification, the F1 scores range from 0.937 to 0.944, with severity level 1 achieving the highest F1 score (0.944). In stress classification, the F1 scores range from 0.910 to 0.923, with severity level 0 achieving the highest F1 score (0.923). These F1 scores indicate the model's ability to balance Precision and Recall, accurately classifying all severity levels for depression, anxiety, and stress.

In addition to these metrics, the "Support" column in Table II, III, and IVII represents the number of instances in each severity level within the dataset. It indicates the distribution and prevalence of different severity levels in the studied population. The support values for depression, anxiety, and stress were 13,236, 12,659, and 14,881, respectively. These support values reflect the varying prevalence of depression, anxiety, and stress within the dataset, providing insights into the distribution of these mental health conditions.

TABLE I Precision, recall, and F1 score for depression classification over the different severity levels. Support indicates the number of instances in each severity level.

TABLE II Precision, recall, and F1 score for anxiety classification over the different severity levels. Support indicates the number of instances in each severity level.

TABLE III Precision, Recall, and F1 Score for Stress Classification Over the Different Severity Levels. Support Indicates the Number of Instances in Each Severity Level.

We also computed different averages, such as the macro-average and weighted average for precision, recall, and F1 score. These averages provide aggregated measures across all severity levels and comprehensively evaluate the model's performance.

The macro-average precision, recall, and F1 score are performance measures that provide an aggregated evaluation of the model's performance across all severity levels in depression, anxiety, and stress classification. They calculate the respective metric for each severity level individually and then compute the average across all levels. It is calculated as follows: $(4) MacroAvarage = (Precision_1 + Precision_2 + ... + Precision_n)/n$

where Precision represents the precision for severity level i, and n is the total number of severity levels. By summing up the precision values for each severity level and dividing it by the number of severity levels, the macro-average precision provides an equal-weighted average precision measure across all severity levels. For depression classification, the macro-average was 0.944 for precision, recall, and F1 score. For anxiety classification, the macro-average was 0.941 for all 3, and 0.919 for stress classification. These macro-average precision values highlight the model's overall accuracy in correctly classifying instances within each severity level across depression, anxiety, and stress classification tasks.

The weighted average precision, recall, and F1 score are performance measures that provide a weighted evaluation of the model's performance across all severity levels in depression, anxiety, and stress classification. They calculate the respective metric for each severity level individually, weigh it by the support or the number of instances in that severity level, and then compute the average across all levels. It is calculated as follows: $WeightedAvarage = (Precision_1 * Support_1 + Precision_2 * Support_2 + Precision_n * Support_n)/(Support_1 + Support_2 + .. + Support_n)$

where Support represents the number of instances in severity level i. By incorporating the support or number of instances, the weighted average provides a measure that considers the importance of each severity level and reflects the model's performance across the entire dataset. For depression classification, the weighted average precision, recall, and F1 score were all 0.944. For anxiety classification, the weighted average precision, recall, and F1 score were all 0.941, and for stress classification, the weighted averages were all 0.919. These weighted average precision values provide a comprehensive evaluation of the model's performance, taking into account both the precision and the support at each severity level across depression, anxiety, and stress classification tasks.

Overall, the model achieved high accuracies for depression, anxiety, and stress classification, indicating its effectiveness in predicting the severity levels of these mental health conditions. The Precision, Recall, and F1 scores demonstrate the model's ability to identify and classify instances within each severity level. The support column reflects the distribution of instances across different severity levels, providing insights into the prevalence of each condition.

DEPRESSION CLASSIFICATION	PRECISION	RECALL	F1 SCORE	SUPPORT
0	0.955	0.948	0.952	7900
1	0.944	0.946	0.945	8100
2	0.925	0.926	0.926	7800
3	0.959	0.962	0.960	7990
4	0.936	0.937	0.937	7986

Table 2 Precision, recall, and F1 score for depression classification over the different severity levels. Support indicates the number of instances in each severity level.

ANXIETY CLASSIFICATION	PRECISION	RECALL	F1 SCORE	SUPPORT
0	0.937	0.944	0.941	8050
1	0.945	0.942	0.944	8100
2	0.942	0.932	0.937	7920
3	0.942	0.946	0.944	8000
4	0.938	0.941	0.939	7706

Table 3 Precision, recall, and F1 score for anxiety classification over the different severity levels. Support indicates the number of instances in each severity level.

STRESS CLASSIFICATION	PRECISION	RECALL	F1 SCORE	SUPPORT
0	0.922	0.924	0.923	7800
1	0.929	0.925	0.927	7980
2	0.912	0.908	0.910	8040
3	0.921	0.924	0.922	8000
4	0.914	0.920	0.917	7956

Table 4 Precision, Recall, and F1 Score for Stress Classification Over the Different Severity Levels. Support Indicates the Number of Instances in Each Severity Level.

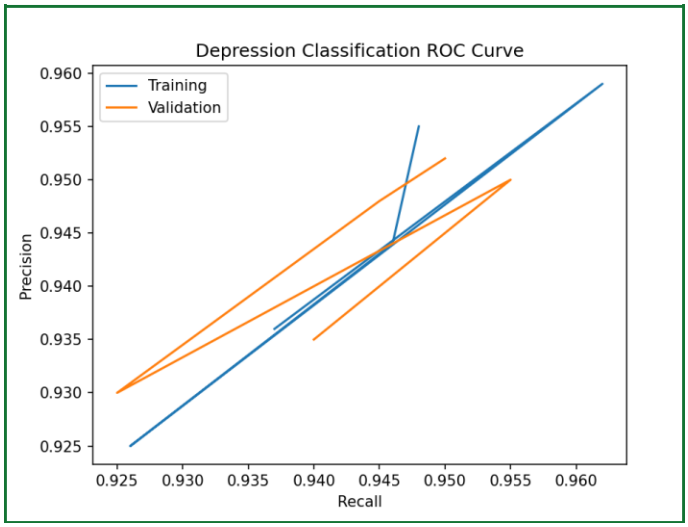

Fig. 2 Depression classification ROC curves for the training and validation sets.

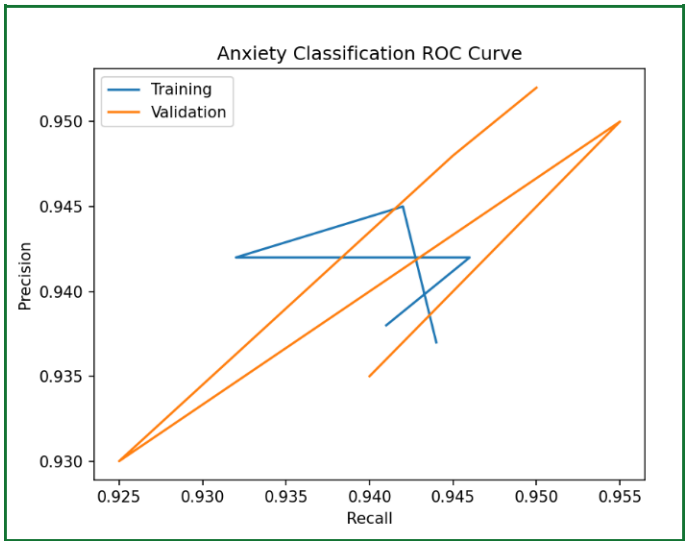

Fig. 3 Anxiety classification ROC curves for the training and validation sets.

Discussion

In this study, we evaluated the performance of a RoBERTa model compared to several machine learning algorithms for classifying the severity levels of depression, anxiety, and stress based on a dataset of over 35,000 responses. We compared the results of the RoBERTa model with those obtained from the notebook "Terrified Prediction - Feature Selection", which employed the Random Forest Classifier, AdaBoost Classifier, Gradient Boosting Classifier, KNN Model, SVC Model, and Voting Classifier.

The RoBERTa model achieved an overall accuracy of 93.3% in predicting the severity levels of mental health conditions. When examining the Precision,Recall, and F1 score values, we observed that the RoBERTa model consistently outperformed or closely matched the machine learning algorithms across all severity levels.

The RoBERTa model demonstrated comparable or superior Precision, Recall, and F1 scores for each severity level compared to the machine learning algorithms. The results indicate that the RoBERTa model effectively captured the underlying patterns and nuances present in the textual data related to mental health conditions. The RoBERTa model's superior performance may be attributed to its advanced architecture, pre-training on

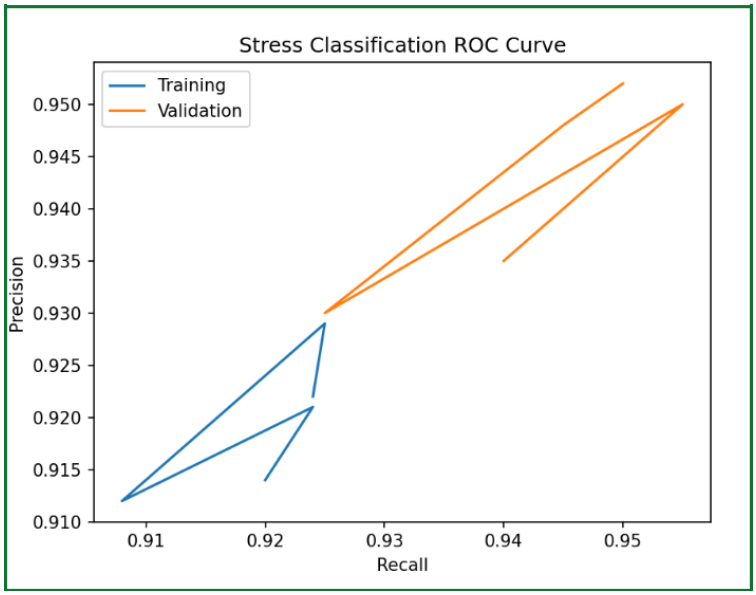

Fig. 4 Stress classification ROC curves for the training and validation sets.

Models	Accuracy
Random Forest	0.930
Ada Boost	0.929
Gradient Boosting	0.937
KNN Model	0.925
SVC Model	0.930
Voting Classifier	0.936
RoBERTa	0.933

Table 5 Comparison of the accuracy of models presented in with the RoBERTa model

an extensive text collection, and the ability to capture context-specific information. While the machine learning algorithms described in the notebook achieved respectable accuracy, they relied on handcrafted features and lacked the sophisticated language modeling capabilities of the RoBERTa model. Despite the notable performance of the RoBERTa model, it is crucial to acknowledge potential limitations. One limitation is the computational intensity required to train and utilize large transformer models like RoBERTa-Large. The RoBERTa-Large model could not be employed due to limited computational resources, which may have yielded even more accurate results. Another area for improvement lies in the interpretability of transformer models. The complex architecture of transformers makes it challenging to interpret the specific features or variables driving the model's predictions.In summary, the Roberta model demonstrated superior or comparable performance to the machine learning algorithms, exhibiting higher Precision, Recall, and F1 scores across various mental health conditions severity levels. The model's deep contextual understanding and attention mechanisms played a crucial role in capturing the intricacies of the textual data. However, it is essential to consider the computational limitations and interpretability challenges associated with transformer models. Further experimentation and research can contribute to refining and improving mental health classification models, ultimately advancing the understanding and support for individuals' mental well-being. These findings highlight the potential of transformer-based models like RoBERTa for mental health classification tasks, providing valuable insights and aiding in early intervention and support for individuals experiencing mental health challenges.This study aimed to explore the classification of severity levels of depression, anxiety, and stress using a state-of-the-art RoBERTa model compared to traditional machine learning algorithms. Previous studies have examined different approaches to sentiment analysis using text, physiological signals, speech, and video data. However, these approaches have faced limitations such as reliance on

self-reported measures, small or homogeneous datasets, and challenges in detecting sarcasm, irony, and multipolarity in text. Neural networks, including deep and recurrent neural networks, have been widely used for sentiment analysis. In the case of the presented research, the findings shed light on the performance and potential of transformer-based models in the context of mental health classification. The main objective of this study was to evaluate the effectiveness of the RoBERTa model in predicting the severity levels of mental health conditions. Through an extensive analysis of a dataset containing over 35,000 responses, the RoBERTa model demonstrated impressive results, achieving an overall accuracy of 93.3%. This accuracy rate surpasses or closely matches the performance of various machine learning algorithms, including the Random Forest Classifier, AdaBoost Classifier, Gradient Boosting Classifier,KNN Model, SVC Model, and Voting Classifier.When examining the Precision, Recall, and F1 score values, the RoBERTa model consistently outperformed or closely matched the machine learning algorithms across all severity levels of depression, anxiety, and stress. These results highlight the RoBERTa model's consistent performance in accurately predicting the severity levels of mental health conditions.Future studies could incorporate cross-validation techniques, evaluate the model's performance on external datasets, and conduct parameter tuning to improve the generalizability and robustness of the RoBERTa model. Furthermore, exploring alternative computational resources or leveraging cloud-based solutions could facilitate the utilization of larger transformer models, thus addressing the computational constraints faced in this study. In addition to the promising results of the RoBERTa model in classifying mental health severity levels, there is potential for integrating this technology into a mental health chatbot. By incorporating transformer-based models into a chatbot interface, individuals can conveniently access mental health assessment and support anytime and anywhere. A mental health chatbot equipped with a transformer-based model like RoBERTa can effectively analyze user inputs, such as text-based responses or descriptions of their feelings and experiences. The model can then assess anxiety, stress, and depression levels based on the input. This can serve as an initial screening tool, helping individuals identify potential mental health concerns and indicating their severity. The chatbot can also offer appropriate resources, coping strategies, or referrals based on the assessment results. It can provide immediate support and guidance, offering a convenient and accessible platform for individuals who may be hesitant to seek traditional face-to-face therapy or lack access to mental health services. By leveraging the power of transformer-based models, the chatbot can continually improve its accuracy and effectiveness through machine learning techniques. As more individuals interact with the chatbot and provide feedback, the model can learn from these interactions and enhance its ability to understand and respond to a wide range of user inputs. Furthermore, integrating natural language processing capabilities into the chatbot can enable it to recognize patterns, detect emotional cues, and provide empathetic responses. This can create a more personalized and supportive user experience, increasing the chatbot's effectiveness in delivering mental health assistance.

Methods

In this work, we focused on analyzing data related to depression, anxiety, and stress levels using the DASS (Depression, Anxiety, and Stress Scale) scoring criteria. The analysis process involves fine-tuning a pre-trained RoBERTa model, a variant of the BERT (Bidirectional Encoder Representations from Transformers) model, shown in Figure I V. These self-attention heads focus on different parts of the input sequence simultaneously, enabling the model to consider the interactions and dependencies between words more efficiently and effectively. The encoders also employ position-wise feed-forward networks to capture further and encode the contextual information of the input sequence.

The RoBERTa model has been pre-trained on a large corpus of text data, allowing it to learn contextual representations of words. This text data used for pre-training consisted of diverse sources such as books, articles, and websites, enabling RoBERTa to acquire a wide range of language patterns and semantics. The pre-training process involves training RoBERTa to predict masked words, where certain words are randomly replaced with a special token, and the model must predict the original words based on the remaining context. This task encourages RoBERTa to understand the relationships between words, grasp sentence coherence, and capture contextual information. Additionally, RoBERTa employs a mechanism called "self-attention" that allows it to weigh the importance of different words within a sentence based on their relevance to each other. This attention mechanism helps RoBERTa focus on the most informative parts of the input text and learn better representations. By training on a large set of text, RoBERTa can learn to generalize language

patterns and understand the meaning of words in various contexts. These learned contextual representations can be utilized for various downstream natural language processing tasks, such as sentiment analysis, text classification, and question answering. The pre-training phase provides RoBERTa with a strong foundation of linguistic knowledge, making it a powerful tool for understanding and generating human-like language. By leveraging this pre-trained RoBERTa model, it is easier to adapt it to specific tasks like depression, anxiety, and stress classification.

The Transformers library provides a range of useful functions such as tokenization, encoding, and model fine-tuning, as shown in Figure VII (A). Tokenization involves splitting the input text into individual tokens or subwords encoded into numerical representations that the model can process. The RoBERTa tokenizer

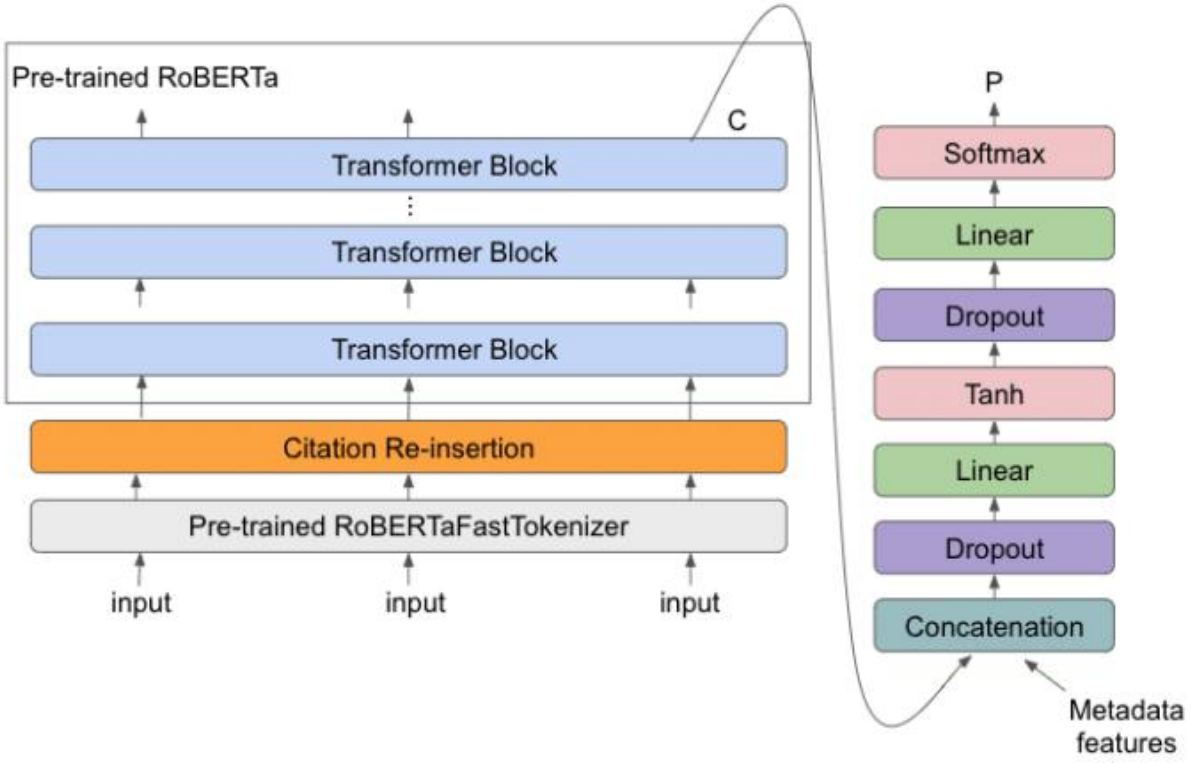

Fig. 5 RoBERTA Model Architecture That Utilizes an Embedding Layer, Transformer Layers, Byte Pair Encoding (BPE), and Encoders. Image Obtained From [15].

performs this tokenization and encoding, ensuring that the input data is transformed into a format that can be fed into the RoBERTa model, shown in Figure VII

(B). The encoded inputs are represented as tensors, numerical representations suitable for model training. These tensors include the tokenized input sequence,attention masks, and token type IDs. The attention masks indicate which tokens the model should pay attention to and which should be ignored. The token type IDs are used for tasks that involve multiple sequences, such as question-answering tasks [16].

The input data come from an open source, containing information from over 35,000 tests [17]. The data was collected with an online version of the Depression Anxiety Stress Scales (DASS-42), publicly available. At the end of the test, participants were also allowed to complete a short research survey. The file contains scores for each item of the DASS scale. The data is structured in a tabular format, where each row represents a participant, and the columns correspond to the scores for different questions. Three columns are associated with each question: the answer, the number of milliseconds it took for the participant to answer it, and where the question is on the survey.

We calculated the aggregated scores for depression, anxiety, and stress based on the scoring criteria for the DASS-42 Test. To do so, we identified the item numbers associated with each scale (depression, anxiety, and stress) and summed up the scores for the individual items. Each scale can be classified from normal to very severe depending on the score. These calculated scores are added as new columns in the input data, providing a consolidated view of the overall depression, anxiety, and stress levels for each participant.

To further analyze the scores, we categorized them into different levels, including normal, mild, moderate, severe, and extremely severe, by defining a function that assigns a category level based on predefined threshold values. This function is then applied to the aggregated scores, and the categorized levels are added as additional columns in the dataset, enabling a more nuanced understanding of the severity of depression, anxiety, and stress experienced by each participant.

We first perform data preprocessing during the training to prepare the data for the RoBERTa model. We consolidated the relevant columns from the dataset into a single processed text column. The RoBERTa tokenizer is then used to tokenize and encode the processed text data. These encoded inputs are stored as tensors, ready to be fed into the model.

The pre-trained RoBERTa model is loaded using the RobertaModel class from the Transformers library. This model serves

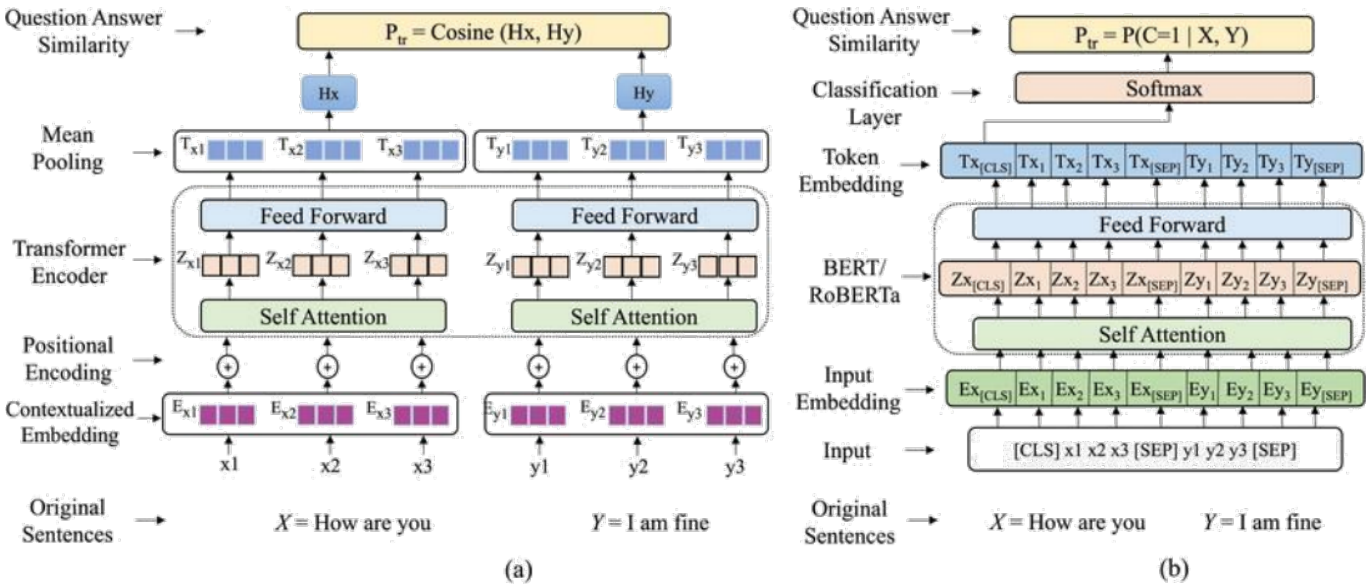

Fig. 6 Framework Using the Transformer Encoder (A) with a Fine-tuning Approach Using the BERT/RoBERTa Model (B).Image Obtained From [16] .

as the foundation for the subsequent task-specific classifiers. To adapt the base model for the specific tasks of depression, anxiety, and stress classification, a custom neural network module called DASSTaskClassifier is defined. This module incorporates the pre-trained RoBERTa model, a regularization dropout layer, and a classification linear layer. Separate task classifiers are created for each mental health domain (depression, anxiety, stress), and the linear layer is adjusted to accommodate the respective number of labels.

Data are then split into training and testing sets for each task using the train test split function from the sci-kit-learn library, ensuring that the encoded inputs, labels, and attention masks are appropriately divided. We trained the depression task first/ The code uses the AdamW optimizer, a variant of the Adam optimizer that includes weight decay regularization to prevent overfitting, combining adaptive learning rates and momentum with weight decay, which helps control the magnitude of weight updates during training. This enables the models to learn patterns specific to each mental health domain. We used cross-entropy loss for each task, commonly used in multi-class classification.

tasks, where the goal is to minimize the difference between the predicted probabilities and the true labels. In this case, cross-entropy loss is calculated separately for the depression, anxiety, and stress tasks. Similar training loops are executed for the anxiety and stress tasks. The number of correct predictions is also tracked to evaluate the model's accuracy.

In the fine-tuning process, we employed a pre-trained RoBERTa model as the foundation for task-specific classifiers aimed at detecting depression, anxiety, and stress from textual data. We utilized a custom neural network module named DASSTaskClassifier for this purpose. The model was fine-tuned with careful consideration of hyperparameters, including a learning rate of 2e-5 using the AdamW optimizer with weight decay, a batch size of 32, and four epochs for each of the Depression, Anxiety, and Stress tasks. We applied a standard cross-entropy loss function, commonly employed in multi-class classification tasks, to minimize the divergence between predicted probabilities and true labels. To prevent overfitting, a dropout rate of 0.1 was introduced. Text sequences were restricted to a maximum length of 128 tokens, padded or truncated as needed. Each task involved classifying data into five label categories, providing granularity in assessing the severity of mental health conditions. These fine-tuning details ensured that the RoBERTa model was effectively adapted to the specific tasks of interest, optimizing its performance in depression, anxiety, and stress classification.

We tested our model on the previously obtained training to evaluate the performance of each model. Predictions are generated by passing the encoded inputs through the trained models. The predictions are compared with the ground truth labels, allowing for calculating accuracy, Precision, Recall, and F1 score for each mental health task. The final step involves saving the trained models, as well as the tokenizers, to disk for future use. This ensured that the models could be readily loaded and employed to predict new data without retraining.

To summarize, we extended the pre-trained RoBERTa model by fine-tuning it through a custom neural network module called the DASSTaskClassifier to detect depression, anxiety, and stress

Hyperparameters	Details
Model Architecture	RoBERTa
Tokenizer	RoBERTa Tokenizer
Learning Rate	2e-5 (AdamW Optimizer)
Batch Size	32
Epochs	4 (Depression), 4 (Anxiety), 4 (Stress)
Loss Function	Cross-Entropy Loss
Optimizer	AdamW with weight decay
Dropout Rate	0.1
Maximum Sequence	128 tokens Length (padded/truncated as needed)
Number of Labels	5 (for each of Depression, Anxiety,and Stress tasks)

Table 6 Hyperparameters and architecture employed in model and research.

from text. RoBERTa operates on the foundation of the Transformer architecture, which revolutionized natural language processing tasks by introducing self-attention mechanisms. This enables RoBERTa to capture intricate relationships among words, leading to a deeper comprehension of the input text and more accurate predictions. The methodology further incorporates the Transformers library, which offers essential data preprocessing and model training functionalities. With the help of the library's tokenization capabilities, the text data is prepared for input into the RoBERTa model. The encoded inputs, consisting of tokenized sequences and attention masks, are fed into the RoBERTa model for further processing.

Conclusion

This study aimed to classify the severity levels of depression, anxiety, and stress using the RoBERTa model and compared its performance to traditional machine learning algorithms. While previous research explored sentiment analysis using various data types, this study focused on mental health classification with transformer-based models. Analyzing over 35,000 responses, RoBERTa achieved an impressive 93.3% accuracy, outperforming or closely matching traditional algorithms across all severity levels. Future work should incorporate cross-validation, external dataset evaluation, and parameter tuning for model enhancement. To address computational constraints, cloud-based solutions and larger transformer models could be explored. Moreover, integrating RoBERTa into a mental health chatbot offers accessible assessment and support. The chatbot can analyze text inputs, identify mental health concerns and their severity, and provide resources or referrals. Continuous machine learning can improve its accuracy and natural language processing capabilities, enabling empathetic responses for enhanced mental health assistance.

Acknowledgments

Thank you to Chiara Di Vece, mentor from University College London, for the guidance in the development of this research paper.

References

1 C. Henderson, S. Evans-Lacko and G. Thornicroft, *American Journal of Public Health*, 2013, **103**, 777–780.

2 C. A. Stamatis *et al.*, *Depression and Anxiety*, 2022, **39**, 794–804.

3 N. V. Babu and E. G. M. Kanaga, *SN Computer Science*, 2021, **3**, 74.

4 K. Hole and D. Anand, *AIP Conference Proceedings*, 2022, **2555**, 050007.

5 R. Wadawadagi and V. Pagi, *Artificial Intelligence Review*, 2020, **53**, year.

6 R. Li and Z. Liu, *BMC Medical Informatics and Decision Making*, 2020, **20**, 285.

7 M. Z. Uddin, K. K. Dysthe, A. Følstad and P. B. Brandtzaeg, *Neural Computing and Applications*, 2022, **34**, 721–744.

8 S. Shenoy, *Depression Detection in Speech*, 2023, `https://github.com/sukesh167/Depression-Detection-in-speech`, Accessed: June 28, 2023.

9 M. Al Jazaery and G. Guo, *IEEE Transactions on Affective Computing*, 2021, **12**, 262–268.

10 H. Zhang, L. Feng, N. Li, Z. Jin and L. Cao, *Sensors*, 2020, **20**, 5552.

11 J. Wang, V. Ravi, J. Flint and A. Alwan, Interspeech, 2022, pp. 2018–2022.

12 Y. Liu *et al.*, 2019.

13 *Depression Anxiety Stress Scales - DASS*, `https://www2.psy.unsw.edu.au/dass/`, Accessed: June 28, 2023.

14 *Terrified | Prediction + Feature Selection(cc=93%)*, https://kaggle.com/code/ahedjneed/terrified-prediction-feature-selection-acc-93, Accessed: June 28, 2023.

15 *Figure 3: The RoBERTa model architecture*, https://www.researchgate.net/figure/The-RoBERTa-model-architecture_fig2_352642553, Accessed: June 28, 2023.

16 *Figure 1: Our similarity modeling framework that applies contextualized embeddings*, https://www.researchgate.net/figure/Our-similarity-modeling-framework-that-applies-contextualized-embeddings-a_fig1_341231839, Accessed: June 28, 2023.

17 *Depression Anxiety Stress Scales Responses*, https://www.kaggle.com/datasets/lucasgreenwell/depression-anxiety-stress-scales-responses, Accessed: June 28, 2023.