IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 January 2024, accepted 12 February 2024, date of publication 16 February 2024, date of current version 26 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3366653

RESEARCH ARTICLE

Harnessing the Power of Hugging Face Transformers for Predicting Mental Health Disorders in Social Networks

ALIREZA POURKEYVAN[1], RAMIN SAFA[1], AND ALI SOROURKHAH[2]
[1]Department of Computer Engineering, Ayandegan Institute of Higher Education, Tonekabon 4681853617, Iran
[2]Department of Management, Ayandegan Institute of Higher Education, Tonekabon 4681853617, Iran

Corresponding author: Ramin Safa (safa@aihe.ac.ir)

ABSTRACT Early diagnosis of mental disorders and intervention can facilitate the prevention of severe injuries and the improvement of treatment results. This study uses social media and pre-trained language models to explore how user-generated data can predict mental disorder symptoms. Our study compares four different BERT models of Hugging Face with standard machine learning techniques used in automatic depression diagnosis in recent literature. The results show that new models outperform the previous approach with an accuracy rate of up to 97%. Analyzing the results while complementing past findings, we find that even tiny amounts of data (Like users' bio descriptions) have the potential to predict mental disorders. We conclude that social media data is an excellent source of mental health screening, and pre-trained models can effectively automate this critical task.

INDEX TERMS Machine learning, mental health, social networks, text mining, transformers.

I. INTRODUCTION

Mental health is undeniably crucial for overall well-being. It impacts an individual's psychological state and has far-reaching effects on physical health [1]. In 2015, the World Health Organization (WHO) reported that approximately 300 million people worldwide suffered from depression. It is estimated that approximately 800,000 people die from depression each year, standing as a major cause of disability globally and contributing to a significant number of suicides and deaths [2].

During the COVID-19 pandemic in 2019, the psychological impact of the virus became increasingly apparent as it spread and people adapted to a new lifestyle. Several studies have found that individuals infected with COVID-19 are more likely to suffer from psychological disorders such as depression [3]. As a result of this situation, mental health issues must be appropriately diagnosed and treated [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Biju Issac.

However, the current diagnosis process for depression faces several limitations. Mental health initiatives face many challenges, including insufficient funding, unequal access to healthcare services, a shortage of mental health specialists, a lack of access to qualified psychologists and psychiatrists, social stigma, and misdiagnosis. Misdiagnosis is a significant concern, as it can impair effective treatment and worsen the condition [1], [5].

In recent years, social media has transformed how people communicate, share information, and connect. A new era of digital communication has been created by the widespread use of social media platforms, which have influenced a number of domains, including marketing, politics, and interpersonal relationships [6], [7]. In turn, social media platforms provide an exceptional opportunity for researchers to gain valuable insights into the mental health of individuals and their professional and social interactions [8], [9].

By leveraging the power of social media and analyzing user-generated content, researchers and healthcare professionals can gain a deeper understanding of mental health

patterns, identify potential risk factors, and develop innovative approaches for diagnosis and treatment. In light of this, researching the relationship between social media use and mental health is an important area of research that may help to improve our understanding of mental illnesses and provide better healthcare outcomes [4].

This paper aims to propose a framework for addressing the limitations of the current diagnostic process and harness the power of transformers and social media to advance our understanding of mental health and develop more effective approaches to supporting individuals experiencing mental illness.

Numerous social network platforms, including Twitter, Facebook, Instagram, and LinkedIn, can be used for data analysis. Millions of active Twitter users make it one of the most popular social media platforms. As a result of its launch in 2006, Twitter has become one of the most popular platforms for real-time news updates, political discussions, and social interaction. According to the latest available statistics, Twitter has more than 330 million monthly active users, making it one of the world's most influential social networks [6]. A machine learning system is constructed using advanced statistical and probabilistic methods to build strategies that can improve through experience. It is a handy tool for predicting mental health. Using it, many researchers can gather information from the data, provide personalized experiences, and develop automated intelligent systems. Support vector machines, random forests, and artificial neural networks are examples of standard methods used in the literature [10].

BERT stands for Bidirectional Encoder Representations from Transformers. In 2018, Google introduced the BERT model, which revolutionized the field of Natural Language Processing (NLP) by performing outstandingly on various language comprehension tasks. As a neural network architecture based on the Transformer model, BERT has a multilayered encoder component based on self-attention mechanisms. The layers contain a feed-forward neural network and a self-attention mechanism. By assigning attention weights based on their importance for understanding the whole sentence, BERT can capture the contextual relationships between different words in a sentence. As a result, BERT can model long-range text dependencies accurately [11], [12].

As part of BERT's training methodology, two key techniques are introduced in addition to Transformer's architecture: masked language modeling (MLM) and next sentence prediction (NSP). By masking out some words randomly during pretraining, BERT is trained to predict the masked words based on the context surrounding the words in a sentence. As a result, BERT can develop deep contextual representations by understanding how words relate to one another. NSP, on the other hand, involves training BERT to determine whether two consecutive sentences appear within a document. By doing so, BERT learns to comprehend the relationships between sentences and grasp the coherence and flow of text by doing so. As a result, BERT can handle tasks such as answering questions and summarising texts, requiring a strong understanding of the relationship between multiple sentences. As part of pretraining, BERT is trained on a large amount of unlabeled text data, using both MLM and NSP objectives. Once these pre-trained representations are in place, they can be fine-tuned for specific downstream tasks, including text classification, question answering, and identifying named entities. In fine-tuning, task-specific layers are applied to the pre-trained BERT model and then trained on labeled data relevant to the specific task [13]. In general, BERT's architecture and the addition of masked language modeling and next-sentence prediction have contributed significantly to advances in NLP. By capturing contextual information, modeling relationships between words and sentences, and applying large-scale pretraining, models can more effectively generate and understand human-like text.

This paper aims to identify the most effective approach to predicting depression from social media data by comparing four different BERT models from Hugging Face with standard machine learning techniques used in literature. We first discuss the recent advances in automated mental health detection from users' social data and then compare several learning models to determine the effectiveness of each in the field to predict depression symptoms. Our results provide insights into the potential of social media for predicting mental states and have implications for early detection and intervention.

This study contributes significantly to mental health monitoring by demonstrating the effectiveness of public social data, specifically bios and tweets, for predicting depression symptoms. While previous research has explored social data for mental health analysis, the specific focus on public information, combined with recent BERT models, represents an unexplored avenue in the literature. The results of our experiments shed light on the untapped potential of public social data as a powerful tool in mental health assessment. Our research sets the stage for future advancements by exploring the untapped potential of social data and emphasizing the importance of comprehensive preprocessing steps, contributing to ongoing efforts in developing mental health diagnosis strategies.

In light of the shortcomings of conventional methods for diagnosing depression and the promising prospects offered by social data, we present the following inquiries for our research: 1) How reliable is the utilization of public social data, particularly user bios and tweets, in accurately forecasting symptoms of depression? 2) When it comes to predicting depression symptoms using public social data, how does the effectiveness of BERT models developed by Hugging Face compare to that of the leading machine learning techniques in the literature? 3) What are the potential ramifications of our discoveries for advancing more efficient strategies for diagnosing and intervening in mental health issues?

The rest of this paper is organized as follows. The next section reviews related work on predicting depression using social data, discussing common approaches, traditional methods, and recent advances in deep learning models like transformers. It also highlights the features studied in pre-

vious works and describes the evaluation metrics used. The third section describes our approach to predicting depression using Twitter data and transformers, providing a detailed design explanation and preprocessing steps. We present our experimental analysis, including implementation details and results. We compare the performance of four BERT models with the cutting-edge machine learning techniques in the literature and discuss implementation challenges. Finally, we discuss the implications of our results, offering insights into the potential of social data for predicting mental health. The paper concludes by summarising our findings and suggesting future research directions.

II. RELATED WORK

To review the latest work, we comprehensively searched relevant keywords in scientific databases, summarised in Table 1. This search obtained a significant number of works in the field of diagnosis of depression and prediction of mental illness, which we will examine the most relevant details in the following.

TABLE 1. Searching strategy of related studies.

No.	Queries
1	Mental Health/ or Mental Disorders / or Electronic Mental Health/ or Depression/ or Suicide/ or Life Satisfaction/ or Well-being/ or Anxiety/ or Stress
2	Social Media/ or Social Networks/ or Twitter/ or Reddit/ or Facebook
3	Machine Learning/ or Data Mining/ or Text Mining/ or Text Analysis/ or Deep Learning / or Language Models/ or BERT/ or Hugging Face/ or Transformers/ or Social Network Analysis/ or Predictive Analysis/ or Prediction/ or Detection
4	(1) and (2)
5	(1) and (3)

Traditionally, most approaches have been based on handcrafted features and rule-based models. Standard methods include sentiment analysis, topic modeling, linguistic analysis/ psycholinguistic analysis, network analysis, and traditional machine learning methods [14]. These tasks will require us to use machine-learning approaches, as shown in Figure 1. We provide a brief overview of the relevant ones in the literature; for more detailed information, we encourage the readers to read references [15] and [16].

Support Vector Machine (SVM) was developed as a popular classification technique. Classification tasks generally involve separating data into training and testing sets. In SVM, the goal is to produce a model based on the training data that predicts, based solely on the test data attributes, the target values of the test data. K-Nearest Neighbor (KNN) is a simple algorithm that stores all the available cases and then classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition as a non-parametric technique.

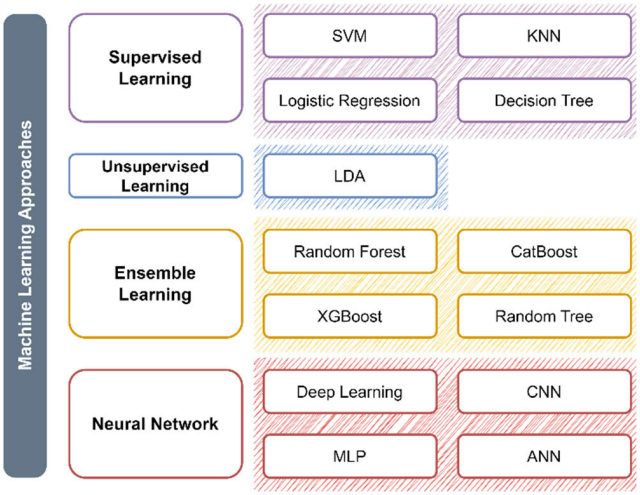


FIGURE 1. Traditional machine learning approaches [18].

Cases are classified based on their neighbors' majority vote and assigned to the class most commonly shared by their K nearest neighbors. If $K = 1$, then the case is simply assigned to the class of its nearest neighbor. The random forest consists of a combination of tree predictors in which every tree is determined by the values of a random vector sampled independently and with the same distribution for all trees within the forest [17].

The Multilayer Perceptron (MLP) is also a type of Artificial Neural Network (ANN) designed to accurately map a given set of inputs to corresponding outputs. It achieves this by combining multiple layers of perceptron, which are mathematical functions, into a single complex function. Each perceptron takes one or more inputs, applies a set of weights and biases to them, and produces an output. The output of each perceptron is then passed through an activation function, which introduces non-linearity into the model, and the resulting output is used as input for the next layer [19]. Its versatility and ability to learn complex relationships make it a popular choice for various applications in various classification fields. It should be noted that most of these approaches have limitations regarding the ability to capture the complex and nuanced language used in social media posts, which is influenced by cultural and contextual factors. A recent development in deep learning models, such as transformers, has enabled researchers to achieve state-of-the-art results in predicting mental health conditions [20].

Transformers are based entirely on a mechanism of attention known as self-attention. An encoder-decoder architecture makes up the transformer. We feed the input sentence to the encoder, which learns the representation of the input sentence and sends the representation to the decoder. The decoder receives input from the encoder, and output is produced. The encoder of the transformer is bidirectional, meaning it can read a sentence both ways. We can perceive BERT as the transformer, but only with the encoder [11]. Recent developments in transformers have shown promising results

when detecting mental health conditions on social media platforms. This architecture has been demonstrated to achieve state-of-the-art performance on various NLP tasks and can capture contextual information. It is also possible to fine-tune transformer models, such as BERT, for specific tasks. Recent studies have demonstrated that BERT effectively captures contextual information and performs at a state-of-the-art level on several NLP tasks [21]. BERT can be fine-tuned for specific tasks, such as depression prediction based on social media data.

Much research has demonstrated that BERT performs better than traditional approaches in detecting mental health conditions on social media platforms [22]. MentalBERT was developed based on posts containing mental health-related information collected from Reddit and is based on BERT-Base. This model follows the standard pretraining protocols for BERT and RoBERTa and is trained and released using Hugging Face's Transformers library to facilitate the automatic identification of mental disorders in online social content using masked language models [23].

Table 2 reviews previous research on depression detection on social media platforms, focusing on traditional approaches such as binary classification and current methods using pre-trained transformer models such as BERT. In conclusion, our review of previous research on depression detection on social media platforms reveals that traditional approaches, such as binary classification, have achieved moderate performance levels. In contrast, new methods utilizing pre-trained transformer models, such as BERT, have demonstrated significantly improved performance.

We will discuss the effectiveness of using pre-trained BERT models from the Hugging Face library to detect mental health issues on social media in the following section, demonstrating that this approach can achieve high accuracy without requiring the creation of new models. By using this approach, we will be able to achieve high performance using fewer computational resources.

TABLE 2. Recent related studies.

Author, date, reference	Mental health issue	Dataset	Platform	Machine learning method(s)
The planned strategy	Depression	Autodep (Textual analysis)	TW	DisitlBERT, BERT, MentalBERT and DistilRoBERTa
Safa et al. (2023) [24]	Depression	Autodep (Multimodal analysis)	TW	Decision Tree, Linear SVM, Gradient Boosting Classifier, Random Forest, RidgeClassifer, AdaBoost, Catboost, and MLP
Kabir et al. (2023) [25]	Depression	Tweets	TW	BERT, DistilBERT
Ilias et al. (2023) [26]	Stress, Depression	Public datasets	RD	MentalBERT
Devika et al. (2023) [27]	Depression, Suicide	Posts	RD	BI-LSTM, BERT
Triantafyllopoulos et al. (2023) [28]	Depression	Pirina dataset	RD	BERT
Benítez-Andrades et al. (2022) [29]	Eating disorder	Tweets	TW	BERT
Zeberga et al. (2022) [5]	Depression, Anxiety	Posts	TW, RD	BI-LSTM, BERT
Vajre et al. (2021) [30]	Mental Health	Conversations	SM	PsychBERT
Nisa et al. (2021) [31]	Depression, Self-harm	Comments	RD	BERT
Bucur et al. (2021) [32]	Gambling, Self-harm, Depression	Posts	RD	BERT
Singh et al. (2021) [33]	COVID-19 impact on social life	Tweets	TW	BERT
Sekulic et al. (2020) [34]	Mental health	SMHD dataset	RD	Logistic Regression, SVM
Cacheda et al. (2019) [35]	Depression	Posts and comments	RD	Random Forest
Peng et al. (2019) [36]	Depression	Bio, Profile, and User's action	TW	SVM
Islam et al. (2018) [37]	Depression	Comments were collected with NCapture	FB	KNN

III. METHODOLOGY

The use of social media in mental health research has gained popularity over the past few years since it provides a rich source of information about users' thoughts, feelings, and behaviors. The potential of social media in mental health research is vast. As previously stated, researchers use it to gain insights into the mental health of a population, track changes in mental health over time, and identify risk and protective factors for mental health issues [38]. Automated systems for detecting mental health issues have been developed to analyze textual data, with NLP and transformer models being utilized to identify patterns [8].

A viable and effective approach for diagnosing mental health disorders within social networks entails examining and analyzing users' self-reported statements. This approach has exhibited promising potential in detecting mental health symptoms, enabling the collection of positive and negative samples that can subsequently be validated automatically to train machine learning models. Previous investigations have demonstrated the efficiency of this method in accurately

identifying mental health symptoms [24], [39]. Nonetheless, the potential impact of BERT models, which have yet to be explored within this specific context, remains uncertain. Given the resounding success of transformers in various domains, our objective is to investigate their effectiveness in the domain of mental health diagnosis using social data.

As part of this study, we propose using Twitter users' tweets and bios to predict depression. Specifically, we use BERT models from the Hugging Face library, which have been fine-tuned based on large datasets of reviews, tweets, and other textual data. A transformer-based deep learning model has demonstrated impressive results in various NLP tasks, such as sentiment analysis, question answering, and text classification [11]. The Hugging Face open-source transformers library in the NLP community is trendy and practical for several Natural Language Understanding (NLU) tasks. The library contains thousands of models that have been pre-trained in more than 100 languages. We can predict symptoms based on users' textual data using pre-trained BERT models from the Hugging Face library. Figure 2 presents a four-step approach that involves data selection, preprocessing, training, and validation modules. To determine which model is best suited to our problem, we tested four BERT models, distilbert-base-uncased-finetuned-sst-2-english (DBUFS2E) [40], bert-base-uncased (BBU) [41], mental-bert-base-uncased (MBBU) [23] and distilroberta-base (DRB) [42]. The following parts describe the methodology in detail, covering fine-tuning and evaluation. Eventually, we analyze the performances of the four models and explore the potential implications for mental health research.

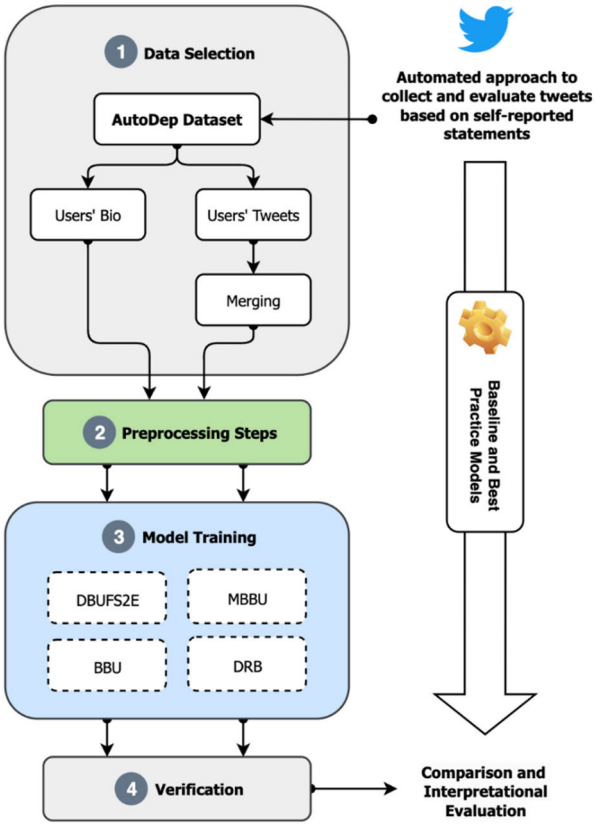


FIGURE 2. Data analysis framework.

A. DATA SELECTION

Data collection is the process of gathering and analyzing information on targeted variables in a systematic manner that enables one to answer stated research questions, test hypotheses, and evaluate results. For training models, data collection is critical in providing the necessary inputs to create models capable of accurately predicting outcomes.

The present study utilizes the Autodep dataset[2], which was automatically collected and evaluated through the Twitter API [24]. It encompasses a range of data, including posts, bio descriptions, profile pictures, and banner images of Twitter users who have publicly disclosed their mental health status. To ensure the authenticity of the results, benchmarking techniques were employed to compare the outcomes of various measures. The dataset contains 11,890,632 tweets and 553 bio-descriptions. In the last study on predictive models for depressive symptoms on the Autodep, utilizing only tweets and bio-texts resulted in accuracy rates of 91% and 83%, respectively [24].

Table 3 and Table 4 display examples of raw tweets and bios in their original form before any preprocessing. We extract the user bios and tweets to work with the data and store them as individual files. To examine the tweets comprehensively, we consolidate all the tweets for each user into a single cell (each tweet is differentiated by a distinct color in the table).

The following section will detail our actions to prepare the data for analysis.

B. PREPROCESSING STEPS

As part of the preparation process for training our model, data preprocessing is an essential step. Social media posts with hashtags, links, special characters, and emojis can often be noisy. However, these elements may not add much value to the text. We used several preprocessing steps to clean and extract meaningful information from the studied features (tweets and bio descriptions).

As illustrated in Algorithm 1, the first step is to check the language, retweets, and mentions. Since they do not add value, they should be skipped. Next, URLs will be eliminated from the text utilizing regular expressions, as URLs are improbable to be associated with the prognosis of mental disorders. URLs have a pattern that begins with http/https or www, and we will use regular expressions to match and remove URLs from the text. Next, we will remove all emojis from the text since they may not provide significant contextual information. We will remove the emojis since we wish to keep our approach as straightforward as possible. Emojis follow an ASCII pattern that can be deleted using a regular expression. We will also remove "∗," "^," "@," etc., using

TABLE 3. Samples of raw tweets.

Tweets	Group
b"I learned that my catatonic phases we're not laziness, but a reaction to having too much on my shoulders.\n\nI had therapy for two years, but it took long until I really could use what I learned in everytime life. I still struggle sometimes.", b'I had a tendency to lie, I wanted to avoid conflict at all cost, and I often fell into phases of almost catatonic disinterest in anything. \nI had therapy to work on my issues, but my doctor did not prescibe antidepressives - something that I now think was a mistake.', b"Today is #WorldMentalHealthDay. Mental health is an important issue that is only going to get more important. \nWhile I can only speak from my own experiences, I think it's important to see that your are not alone in your struggle. So here goes:"	Diagnosed
b'Not inciting anything. Just want the world to know my story. 18,000 plus pieces of content.', b'All I ever had was my Bible and poetry to fight with. All year every year', b'\xf0\x9f\x93\xa3 New Podcast! "Normalised: The One Unforgivable Sin (11/15/20)" on @Spreaker https://t.co/wiesJLZegY"'	Controlled

TABLE 4. Samples of raw bio descriptions.

Bio	Group
Small time streamer, part time musician, full time audio engineer. He/him.	Diagnosed
Prophet of the Living God. Author. A collection of spiritual and societal truths unearthed along a journey called life. âœ• ï¸• #backtrack #christianlivesmatter	Controlled

regular expressions to improve noise reduction and a smaller dataset.

As part of the removal process, we will make sure to eliminate any extra spaces that might remain. Following, the sentence will be converted to lowercase to ensure consistency. As text data can have different case types, the conversion to lowercase will ensure that the same word will be treated as identical regardless of its case. To enhance the quality of the text, we shall remove any extra spaces between tweets that may impact the algorithm's performance.

After that, we will tokenize the text using the NLTK library, which will transform long texts into smaller units known as tokens. Tokenization is the process of splitting the text into individual words or tokens that can then be used as input for further analysis.

We will perform using the Natural Language Toolkit NLTK [43] library to eliminate insignificant and meaningless stop words from the English language, such as 'I,' 'am,' 'a,' 'the,' 'of,' 'to,' etc. And we will apply lemmatization to group different inflected forms of words to allow them to be analyzed together as a single item. Lemmatisation is a process of normalizing words into their base or root form, which reduces the number of variables to reduce the complexity

Algorithm 1 Data Preprocessing Steps

Input:
Tweets and Bio descriptions (control and diagnosed group)
Output:
Preprocessed data
Method:
for each value in groups
if the value is not in English, skip the value
if the value starts with "RT," skip the value
if the value starts with "@," skip the value
if the value contains URLs, remove them from the value
if the value contains emoji, remove them from the value
if the value contains any special character, sanitize the value convert the value to lowercase
if the value contains any additional spaces, remove them tokenize the value
if the value contains stop words, remove them lemmatize values
end for

of text data. A lemmatization will be performed using the WordNetLemmatizer[3] from the NLTK library. We will also be conducting experiments with stemming, which reduces words to their basic form by eliminating suffixes from their root form. However, we found that lemmatization provided slightly better results than stemming, with a minimal difference. After completing the preprocessing steps, we will verify that the length of the character does not fall below five characters to remove any unnecessary words or characters that may have been left over from the previous steps. Then, we will return the results to the training section. Figure 3 provides an example of the aforementioned preprocessing steps as follows.

C. MODEL TRAINING

Upon completion of the preprocessing steps, we will utilize the Hugging Face library to fine-tune four pre-trained BERT models: distilbert-base-uncased-finetuned-sst-2-english, bert-base-uncased, distilroberta-base, and mental-bert-base-uncased. This approach will enable us to develop a predictive model for detecting depression.

Distilbert-base-uncased-finetuned-sst-2-english is a text classification model developed by Hugging Face that uses distilbert-base-uncased for topic classification, which was trained on the SST-2 dataset. In addition to fine-tuning downstream tasks, the model can also be used to model masked language or predict the next sentences. The distilBERT model has been developed to simplify the original BERT model, making it smaller, faster, and more efficient while maintaining most of its performance.

The bert-base-uncased model is a pre-trained NLP model introduced in a paper. This model has been trained on a large corpus of English data using a self-supervised approach; in other words, it has been trained on raw texts without human labeling. A model was trained on two objectives: Masked

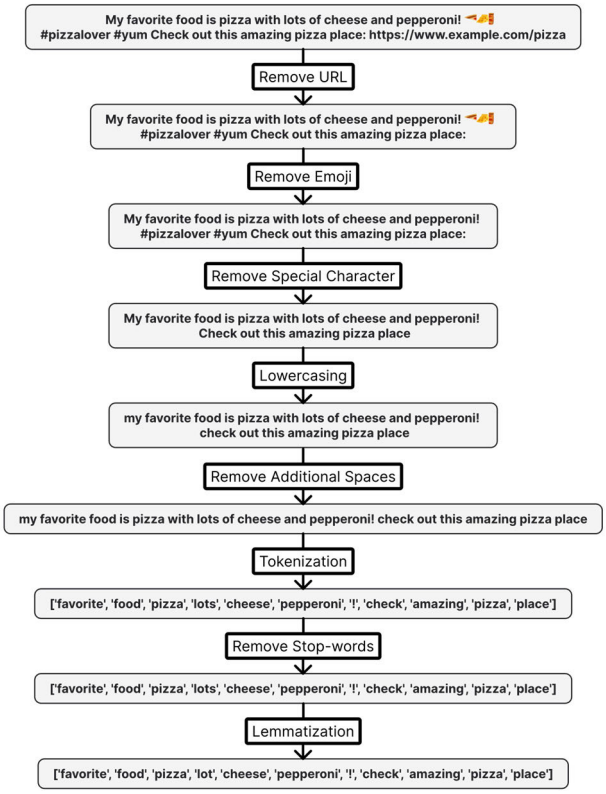


FIGURE 3. Preprocessing steps.

Language Modelling (MLM) and Next Sentence Prediction (NSP). MLM involves randomly masking 15% of the words in a sentence and predicting the masked words, allowing the model to learn a bidirectional sentence representation. In NSP, two masked sentences are concatenated, and a prediction is made about whether the two sentences follow each other. The learned features of the BERT model can be applied to downstream tasks, such as classification.

The Roberta-base model was pre-trained on a large corpus of English data using the MLM objective without human labeling. The model can learn a bidirectional sentence representation to extract useful features for downstream tasks, such as sequence classification, token classification, or question answering. It is case-sensitive, meaning it differentiates between different capitalizations. There is a distilled version of the Roberta-base model, distilroberta-base, which follows the same training procedure as DistilBERT [44]. Compared to the Roberta-base model, it has fewer layers, dimensions, and parameters, resulting in a smaller, faster, and more efficient model while maintaining its performance in most cases.

As discussed earlier, MentalBERT is a variant of the bert-base-uncased model that has been fine-tuned on a dataset of mental health-related posts from Reddit. It allows the model to capture better the nuances and complexities of mental health language, which can be useful for tasks such as sentiment analysis, classification, or question-answering on related content. The learned features of the BERT model can be applied to downstream tasks, and the use of MentalBERT represents a significant step forward in using NLP to address mental health issues and improve our understanding of this crucial area. We found that utilizing pre-trained models gave our approach a significant advantage. Using pre-existing weights and architectures decreased the computational resources needed for training while enhancing the model's performance. As a result of their training on large quantities of text data, these models could learn complex patterns and relationships within the text, which is crucial for detecting depression.

As a part of our methodology, we loaded the datasets as CSV files and divided them into training and testing sets using an 80/20 split. It is worth noting that two distinct datasets were used to apply this split: tweets and bios. As a next step, we applied tokenization to the text data using the pre-trained BERT tokenizer specific to the model we were fine-tuning. These tokens are then utilized as inputs in the model for further analysis and processing.

We utilized the Hugging Face library and the Trainer method to load, preprocess, and fine-tune the pre-trained BERT models. To load the pre-trained BERT model with the specified number of labels (in this case, two for binary classification), we used the AutoModelForSequenceClassification class.

As an integral component of the training process, we partitioned the datasets into ten distinct folds and shuffled them before initiating the training process. Using K-fold cross-validation, we avoided overfitting and improved model accuracy. Each fold was trained separately, using five epochs per training. The model was trained on the training set and evaluated on the validation set at each training epoch. We loaded, preprocessed, and fine-tuned pre-trained BERT models for both tweets and bios using the Hugging Face library and the Trainer method. As a result of fine-tuning the pre-trained models with our preprocessed Twitter dataset, we could construct high-performing depression detection models. The K-fold cross-validation technique ensures that our model is trained and evaluated on a diverse dataset, thus ensuring quality and accuracy.

An evaluation function was used to compute various evaluation metrics during the training process, including accuracy, F1 score, and AUC, to provide insight into the quality and performance of the model in question. To construct high-performance depression detection models, we used pre-trained BERT models that were fine-tuned based on preprocessed Twitter datasets. By applying K-fold cross-validation, we were able to train and evaluate our models on a variety of datasets, thereby ensuring that our results would be reliable and accurate. The partial implementation of this process can be found in our GitHub repository[1].

D. VALIDATION

Several metrics were used to evaluate the performance of our fine-tuned BERT models for depression detection, including accuracy, F1 score, receiver operating characteristic (ROC), and the area under the curve (AUC).

An accuracy measure indicates how many instances were correctly predicted. The F1 score measures the balance between precision and recall and represents the harmonic mean of precision and recall. Precision measures the proportion of true positive predictions among all positive predictions made by the model. Conversely, Recall measures the proportion of true positive predictions among all actual positive instances in the data. AUC indicates how well the model performed regarding true positive rates (TPRs) and false positive rates (FPRs). An AUC represents the area under a ROC curve, which plots the TPR against the FPR. By using the confusion matrix derived from the predictions of the models, the accuracy and F1 score were calculated using Equations 1 to 4 for each fold:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{1}$$

To calculate the F1 score, we first compute the precision and recall using Equation 2 and Equation 3.

$$Precision = \frac{TP}{(TP + FP)}) \tag{2}$$

$$Recall = \frac{TP}{(TP + FN)}) \tag{3}$$

$$F1score = \frac{2(precision \times recall)}{(precision + recall)}) \tag{4}$$

TP (True Positive) represents the number of correctly predicted positive instances, TN (True Negative) represents the number of correctly predicted negative instances, FP (False Positive) represents the number of incorrectly predicted positive instances, and FN (False Negative) represents the number of incorrectly predicted negative instances.

For the final evaluation of the performance of the models, we used K-fold cross-validation. We split the dataset into ten folds and trained each fold separately, using five epochs per training. At each epoch, the model was trained on the training set and evaluated on the validation set.

To assess the final performance of the models, we employed K-fold cross-validation. Specifically, we partitioned the dataset into ten folds and trained each fold individually, with a training duration of five epochs. At each epoch, the model was trained on the training set and evaluated on the validation set. This process was repeated for each tenfold, resulting in a comprehensive evaluation of the model's performance across the entire dataset. By utilizing K-fold cross-validation, we aim to provide a robust and reliable assessment of the model's performance while minimizing the risk of overfitting to a specific subset of the data.

IV. EXPERIMENTAL ANALYSIS

In the previous sections, we discussed various concepts that provide the foundation for our methodology for predicting depression. This section examines experimental analysis in greater depth to assess our approach's effectiveness.

Our implementation was carried out using Python programming language and the Google Colab platform, which enabled us to leverage the computing power of a T4 GPU. We utilized two distinct datasets comprising users' tweets and bios to fine-tune the pre-trained BERT models. This process involved adapting the pre-existing models to our data's specific domain, which helped optimize their performance on our task. By fine-tuning the BERT models on these datasets, we could tailor their representations to capture the unique characteristics of our data and achieve state-of-the-art results.

TABLE 5. Accuracy, F1 score, and AUC for tweets analysis.

Model	Accuracy	F1 score	AUC
DBUFS2E	0.97	0.96	0.98
BBU	0.97	0.97	0.98
MBBU	0.96	0.96	0.98
DRB	0.95	0.95	0.97
Baseline	0.75	0.75	0.75
Catboost	0.91	0.89	0.91
GBC	0.91	0.89	0.90

TABLE 6. Accuracy, F1 score, and AUC for bios analysis.

Model	Accuracy	F1 score	AUC
DBUFS2E	0.95	0.96	0.96
BBU	0.95	0.94	0.96
MBBU	0.96	0.96	0.96
DRB	0.95	0.95	0.96
Baseline	0.67	0.62	0.67
MLP	0.83	0.82	0.83
RidgeClassifier	0.71	0.68	0.71

To further strengthen the robustness of our assessment and provide a comprehensive evaluation, we compared our approaches with the baseline (Logistic Regression) and four leading machine learning models in the prior study [24]. This study used standard evaluation metrics, such as accuracy, F1 score, and AUC, to compare multiple models to predict depression and related disorders. According to the investigation, it was consistently observed that Catboost, Gradient Boosting Classifier (GBC), MLP, and RidgeClassifier consistently outperformed other approaches regarding the evaluated metrics. Consequently, we opted for these models and proceeded to implement our methods on the same datasets, enabling us to evaluate the aforementioned metrics comprehensively.

The results of our analysis are presented in Tables 5 and 6. As shown, our approach achieved competitive performance compared to previous models. With the help of our models, DBUFS2E, BBU, MBBU, and DRB, we consistently outperformed the previous techniques in terms of accuracy, F1 score, and AUC across both datasets. The superior performance of our models over fully trained BERT models further highlights the effectiveness of our approach and its ability to extract meaningful insights from Twitter data for depression detection. These results demonstrate the potential of fine-tuning pre-trained language models for specific tasks, particularly in the context of social media-based mental health assessment.

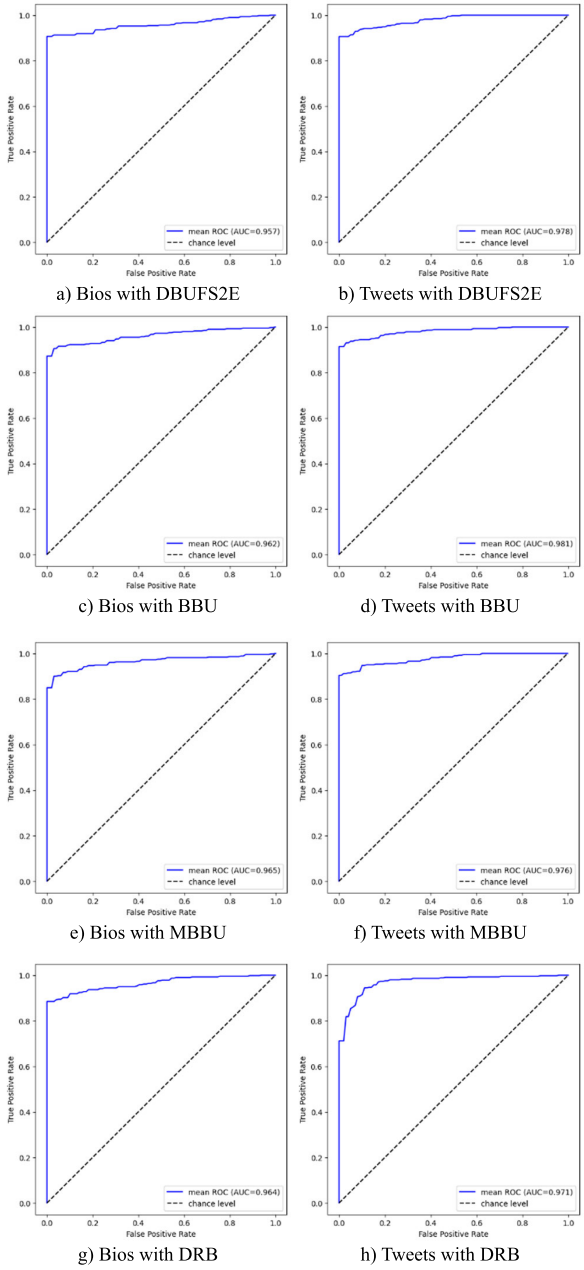


FIGURE 4. ROC curves for bios and tweets.

By including a range of models in our evaluation, we aim to provide a more complete picture of the strengths of each approach, which can help to inform future research in this area. Based on the results above, it is apparent that the DBUFS2E model exhibits the highest metric values among all models for both the Tweet and Bio datasets. Notably, this model significantly outperforms other standard methods used in prior research. These findings corroborate previous studies demonstrating a strong association between user-generated content and mental health status. The short bio-text substantially impacts the detection of clues related to depressive disorder. These results underscore the potential of these models for early detection and prevention of mental health issues.

Figure 4 also shows the ROC curves for DBUFS2E, BBU, MBBU, and DRB models on the tweet and bio datasets. For both datasets, the DBUFS2E model achieved the highest mean AUC values, with the bio dataset achieving a mean AUC of 0.96 and the tweet dataset achieving a mean AUC of 0.98. Additionally, the BBU model performed well, with a mean AUC value of 0.95 and 0.97 for the bio and tweet sets. According to the ROC curves, all models achieved high TPRs and low FPRs, indicating that the models could accurately detect depression. This section explored using four different BERT models for predicting depression. We compared these models to traditional approaches and found that BERT models outperformed traditional approaches in terms of accuracy. One of the key advantages of BERT models is their ability to capture contextual information, considering the entire sentence or document rather than individual words in isolation. It allows them to capture subtle nuances and dependencies within the text, which are crucial for accurate predictions.

Additionally, BERT models benefit from pre-training on large-scale datasets, giving them a wide range of language patterns and semantic relationships. Finally, BERT models excel at handling long-range dependencies in text, capturing the global context and leading to more accurate predictions for depression detection. By harnessing the power of BERT models, we can potentially improve the accuracy and efficacy of depression prediction, enhancing mental health care and support for individuals in need.

V. CONCLUSION AND FUTURE WORK

Our study underscores the importance of analyzing social media data for mental health research, as it can provide valuable insight into an individual's mental health status. By examining Twitter data, we demonstrate the potential of using such information to predict depression symptoms with high accuracy and F1 scores compared to previous findings, which is a promising development in the automatic detection of depression symptoms. We analyzed tweets and bio information using four different BERT models from the Hugging Face and evaluated the models' performance using 10-fold cross-validation and standard measurements in the literature. The results show a more than 6% increase in metrics compared to previous research.

It should be noted that additional user data, including images, emojis, and hashtags, can enhance predictions. Future research can explore combining data types to improve the performance of the models; for example, a more comprehensive view of mental health status can be obtained by combining tweets and bio information or by modifying preprocessing steps so that useful data such as emojis is retained. Emojis may carry valuable emotional cues, and translating them into meaningful representations could provide insights into the emotional state of individuals. By incorporating this additional information, we can examine whether it contributes significant value to the predictive results or enhances the models' performance.

Moreover, another avenue for further investigation would be to examine the performance of BERT models compared to other transformer-based models, such as the GPT model. Using this comparison, we will gain insight into the strengths and weaknesses of various transformer architectures concerning mental health analysis. Additionally, our approach may also be applied to other mental health conditions to improve early detection and intervention, such as anxiety and post-traumatic stress disorder. Exploring the effectiveness of social data mining in these domains would contribute to a broader understanding of mental health and facilitate the development of targeted interventions.

Our research outcomes present compelling evidence that directly addresses the initial research inquiries we posed:

1) The outcomes of our study strongly indicate that public social data can be effectively utilized to predict symptoms of depression.

2) We observe that the performance of BERT models outperforms that of leading machine learning techniques in the literature, underscoring the considerable potential of deep learning in the realm of mental health analysis.

3) The implications of our findings are far-reaching in terms of the early detection and intervention of depression. The capability to forecast depression symptoms by leveraging public social media data can empower individuals, healthcare practitioners, and mental health organizations to identify those at risk and provide timely support.

These findings not only align with our initial research questions but also offer valuable insights into the potential utilization of social media for monitoring mental health and facilitating early intervention.

We hope this study inspires future research into social data mining for mental health research and ultimately improves mental health outcomes for individuals worldwide.

APPENDIX

FOOTNOTES

[1]https://github.com/rsafa/BERT4MentalHealthMonitoring

[2]https://github.com/rsafa/autodep

[3]https://nltk.org/_modules/nltk/stem/wordnet.html

ACKNOWLEDGMENT

The authors would like to acknowledge the developers of hugging face transformers for making their pre-trained language models publicly available.

REFERENCES

[1] M. Prince, V. Patel, S. Saxena, M. Maj, J. Maselko, M. R. Phillips, and A. Rahman, "No health without mental health," *Lancet*, vol. 370, no. 9590, pp. 859–877, Sep. 2007, doi: 10.1016/s0140-6736(07)61238-0.

[2] *Depression and Other Common Mental Disorders: Global Health Estimates*, World Health Organization, Geneva, Switzerland, 2017.

[3] O. Renaud-Charest, L. M. Lui, S. Eskander, F. Ceban, R. Ho, J. D. Di Vincenzo, J. D. Rosenblat, Y. Lee, M. Subramaniapillai, and R. S. McIntyre, "Onset and frequency of depression in post-COVID-19 syndrome: A systematic review," *J. Psychiatric Res.*, vol. 144, pp. 129–137, Dec. 2021.

[4] S. Dhelim, L. Chen, S. K. Das, H. Ning, C. Nugent, G. Leavey, D. Pesch, E. Bantry-White, and D. Burns, "Detecting mental distresses using social behavior analysis in the context of COVID-19: A survey," *ACM Comput. Surv.*, vol. 55, no. 14s, pp. 1–30, Dec. 2023.

[5] K. Zeberga, M. Attique, B. Shah, F. Ali, Y. Z. Jembre, and T.-S. Chung, "A novel text mining approach for mental health prediction using bi-LSTM and BERT model," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–18, Mar. 2022.

[6] Statista. (2021). *Number of Monthly Active Twitter Users Worldwide From 1st Quarter 2010 to 1st Quarter 2021*. [Online]. Available: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

[7] Statista. (Sep. 2019). *Reasons for US Users to Follow Brands on Twitter*. [Online]. Available: https://www.statista.com/statistics/276393/reasons-for-us-users-to-follow-brands-on-twitter/

[8] A. Le Glaz, Y. Haralambous, D.-H. Kim-Dufor, P. Lenca, R. Billot, T. C. Ryan, J. Marsh, J. DeVylder, M. Walter, S. Berrouiguet, and C. Lemey, "Machine learning and natural language processing in mental health: Systematic review," *J. Med. Internet Res.*, vol. 23, no. 5, May 2021, Art. no. e15708.

[9] M. Casavantes, M. E. Aragón, L. C. González, and M. Montes-Y-Gómez, "Leveraging posts 'and authors' metadata to spot several forms of abusive comments in Twitter," *J. Intell. Inf. Syst.*, vol. 61, pp. 1–21, Feb. 2023.

[10] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, Jul. 2015.

[11] S. Ravichandiran, *Getting Started With Google BERT: Build and Train State-of-the-Art Natural Language Processing Models Using BERT*. Birmingham, U.K.: Packt, 2021.

[12] S. M. Jain, "BERT," in *Introduction to Transformers for NLP: With the Hugging Face Library and Models to Solve Problems*. Berkeley, CA, USA: Apress, 2022, pp. 37–49, doi: 10.1007/978-1-4842-8844-3_3.

[13] D. Rothman and A. Gulli, *Transformers for Natural Language Processing: Build, Train, and Fine-Tune Deep Neural Network Architectures for NLP With Python, PyTorch, TensorFlow, BERT, and GPT-3*. Birmingham, U.K.: Packt, 2022.

[14] C. C. Aggarwal, *Machine Learning for Text*, vol. 848. Berlin, Germany: Springer, 2018.

[15] A. Géron, *Hands-on Machine Learning With Scikit-Learn, Keras, and TensorFlow*. Sebastopol, CA, USA: O'Reilly Media, 2022.

[16] A. C. Müller and S. Guido, *Introduction to Machine Learning With Python: A Guide for Data Scientists*. Sebastopol, CA, USA: O'Reilly Media, 2016.

[17] R. A. Nugrahaeni and K. Mutijarsa, "Comparative analysis of machine learning KNN, SVM, and random forests algorithm for facial expression classification," in *Proc. Int. Seminar Appl. Technol. Inf. Commun. (ISemantic)*, Aug. 2016, pp. 163–168.

[18] J. Chung and J. Teo, "Mental health prediction using machine learning: Taxonomy, applications, and challenges," *Appl. Comput. Intell. Soft Comput.*, vol. 2022, pp. 1–19, Jan. 2022.

[19] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[21] R. Pandey and J. P. Singh, "BERT-LSTM model for sarcasm detection in code-mixed social media post," *J. Intell. Inf. Syst.*, vol. 60, no. 1, pp. 235–254, Feb. 2023.

[22] S. González-Carvajal and E. C. Garrido-Merchán, "Comparing BERT against traditional machine learning text classification," 2020, *arXiv:2005.13012*.

[23] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "MentalBERT: Publicly available pretrained language models for mental healthcare," in *Proc. LREC*, 2022, pp. 1–8.

[24] R. Safa, P. Bayat, and L. Moghtader, "Automatic detection of depression symptoms in Twitter using multimodal analysis," *J. Supercomput.*, vol. 78, no. 4, pp. 4709–4744, Mar. 2022.

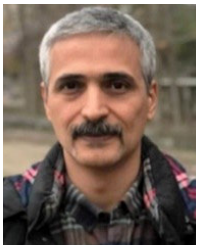
[25] M. Kabir, T. Ahmed, M. B. Hasan, M. T. R. Laskar, T. K. Joarder, H. Mahmud, and K. Hasan, "DEPTWEET: A typology for social media texts to detect depression severities," *Comput. Hum. Behav.*, vol. 139, Feb. 2023, Art. no. 107503.

[26] L. Ilias, S. Mouzakitis, and D. Askounis, "Calibration of transformer-based models for identifying stress and depression in social media," *IEEE Trans. Computat. Social Syst.*, early access, Jun. 16, 2023, doi: 10.1109/TCSS.2023.3283009.

[27] S. P. Devika, M. R. Pooja, M. S. Arpitha, and R. Vinayakumar, "BERT-based approach for suicide and depression identification," in *Proc. 3rd Int. Conf. Adv. Comput. Eng. Commun. Syst.*, 2023, pp. 435–444.
[28] I. Triantafyllopoulos, G. Paraskevopoulos, and A. Potamianos, "Depression detection in social media posts using affective and social norm features," 2023, *arXiv:2303.14279*.
[29] J. A. Benítez-Andrades, J.-M. Alija-Pérez, M.-E. Vidal, R. Pastor-Vargas, and M. T. García-Ordás, "Traditional machine learning models and bidirectional encoder representations from transformer (BERT)—Based automatic classification of tweets about eating disorders: Algorithm development and validation study," *JMIR Med. Informat.*, vol. 10, no. 2, Feb. 2022, Art. no. e34492.
[30] V. Vajre, M. Naylor, U. Kamath, and A. Shehu, "PsychBERT: A mental health language model for social media mental health behavioral analysis," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2021, pp. 1077–1082.
[31] Q. U. Nisa and R. Muhammad, "Towards transfer learning using BERT for early detection of self-harm of social media users," in *Proc. Work. Notes CLEF*, 2021, pp. 21–24.
[32] A.-M. Bucur, A. Cosma, and L. P. Dinu, "Early risk detection of pathological gambling, self-harm and depression using BERT," 2021, *arXiv:2106.16175*.
[33] M. Singh, A. K. Jakhar, and S. Pandey, "Sentiment analysis on the impact of coronavirus in social life using the BERT model," *Social Netw. Anal. Mining*, vol. 11, no. 1, p. 33, 2021.
[34] I. Sekulić and M. Strube, "Adapting deep learning methods for mental health prediction on social media," 2020, *arXiv:2003.07634*.
[35] F. Cacheda, D. Fernandez, F. J. Novoa, and V. Carneiro, "Early detection of depression: Social network analysis and random forest techniques," *J. Med. Internet Res.*, vol. 21, no. 6, Jun. 2019, Art. no. e12554.
[36] Z. Peng, Q. Hu, and J. Dang, "Multi-kernel SVM based depression recognition using social media data," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 1, pp. 43–57, Jan. 2019.
[37] M. R. Islam, A. R. M. Kamal, N. Sultana, R. Islam, M. A. Moni, and A. Ulhaq, "Detecting depression using K-nearest neighbors (KNN) classification technique," in *Proc. Int. Conf. Comput., Commun., Chem., Mater. Electron. Eng. (ICME)*, Feb. 2018, pp. 1–4.
[38] L. Braghieri, R. Levy, and A. Makarin, "Social media and mental health," *Amer. Econ. Rev.*, vol. 112, no. 11, pp. 3660–3693, 2022.
[39] R. Safa, S. A. Edalatpanah, and A. Sorourkhah, "Predicting mental health using social media: A roadmap for future development," 2023, *arXiv:2301.10453*.
[40] *Distilbert-Base-Uncased-Finetuned-SST-2-English (Revision BFDD146)*, HF Canonical Model Maintainers, Hugging Face, New York, NY, USA, 2022, doi: 10.57967/hf/0181.
[41] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[42] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.
[43] S. Bird, "NLTK: The natural language toolkit," in *Proc. COLING/ACL Interact. Presentation Sessions*, 2006, pp. 69–72.
[44] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

ALIREZA POURKEYVAN is currently pursuing the M.Sc. degree in computer engineering with the Ayandegan Institute of Higher Education, Iran. He actively explores innovative approaches to extract insightful information from textual data using transformers and develops intelligent systems for language understanding and analysis. His research interest includes natural language processing (NLP) applications.

RAMIN SAFA received the M.Sc. degree in information technology from the University of Guilan, Iran, in 2014, and the Ph.D. degree in computer engineering from Islamic Azad University, Iran, in 2022. He is currently the Technical Director of the Innovation Hub, Ayandegan Institute of Higher Education, Iran. His research interests include text mining and machine learning applications for health care. As a Reviewer, he collaborates with numerous international journals, such as *The Journal of Super Computing*, *Journal of Medical Internet Research*, and IEEE ACCESS.



ALI SOROURKHAH is currently an Assistant Professor with the Department of Management, Ayandegan Institute of Higher Education, Iran. He published various technical papers in peer-reviewed journals and conference proceedings. His current research interests include soft operational research, decision-making, problem-structuring methods, and statistical analysis.