IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 May 2024, accepted 24 May 2024, date of publication 28 May 2024, date of current version 6 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3406469

RESEARCH ARTICLE

Detection and Prediction of Future Mental Disorder From Social Media Data Using Machine Learning, Ensemble Learning, and Large Language Models

MOHAMMED ABDULLAH[1] AND NERMIN NEGIED[2,3], (Member, IEEE)
[1]School of Electrical and Computer Engineering, University of Ottawa, Ottawa, ON K4A 6N5, Canada
[2]Digital Egypt Builders Initiative, Egyptian Ministry of Communication and Information Technology, Cairo 11531, Egypt
[3]School of Information Systems and Computer Science, Nile University, Giza 12677, Egypt

Corresponding author: Nermin Negied (nnegied@nu.edu.eg)

ABSTRACT Social media platforms are used widely by all people to express their feelings, opinions, and emotional states. Billions of people worldwide use them daily to share what they think and feel in their posts. Amongst all social media available platforms, Facebook only contains around three billion personal accounts. In this work Reddit dataset is used to automatically detect mental illness from social media posts. This study is not only limited to early detection of already existing mental illness or disorder like depression and anxiety from social posts, but also and most importantly the study is extended to predict successfully potential mental illness that would happen in future. This study deploys Nineteen different models to study the capability of them in detecting and predicting mental disorders from social media posts. Some of the deployed models are classical machine learning classifiers, some are ensemble learning models, and the rest are large language models (LLMs). Six machine learning classifiers were used in this work for the automatic detection and prediction of mental illness and logistic regression proved to be the best amongst other classifiers in this task. Nine Ensemble methods were also used and examined. Amongst the Nine ensemble learning models VC2, Light GBM, Bagging estimator, and XGBoost proved to be superior in this task. Four large language models were also used and examined for the same task. RoBERTa and OpenAI GPT proved to outperform the rest of models in this task. All those models were built, trained, tested, and compared with previous work in literature to get the best possible results. The study covers the main four mental disorders which are ADHD, Anxiety, Bipolar, and Depression. The work proposed in this paper succeeded in outperforming the results in literature in terms of number of addressed mental disorders, number of models used and tested, and dataset size used to validate results. The proposed work also outperformed the only attempt in literature that addressed all mental disorders in results of detection and prediction noticeably. This work achieved the detection of already existing mental disorders F1-score of 0.80 from clinical data and of 0.52 from non-clinical data, and it achieved a prediction of future mental disorder F1-score of 0.43 from non-clinical data.

INDEX TERMS ADHD, anxiety, bipolar, mental illness classification, depression detection, depression prediction, Reddit.

I. INTRODUCTION

Social media platforms are nowadays used by almost every single person on earth. People use it to express their feelings and attitudes towards everything, including other people, products, weather, social events, and political issues [1]. Natural language processing (NLP) and new AI techniques have also led to the development of many new technological advancements that are used by all people in their ordinary life.

The associate editor coordinating the review of this manuscript and approving it for publication was Maria Chiara Caschera.

Most popular NLP and AI modern techniques used by people worldwide are Machine translation, information extraction, information retrieval, question answering, text memorization, automatic assistance, and recommendation chat-bots and apps, etc. [2], [3], [4], [5]. Together Social media platforms with NLP and AI have made a lot of things easier to people, like fast communication across different countries, giving people the ability to know news that happens everywhere on spot, not just that but it gave them the ability to express what they think and feel using posts and comments. Those facts led to having huge chunks of data publicly available on the internet. Those huge chunks of data, especially text data have been used by NLP, AI, and Data Science researchers in many purposes including automatic sentiment analysis, network analysis, information extraction, knowledge graph building, information interpretation, etc.

In the last few years, a new direction was added to the pre-mentioned research directions that benefit from data on social media, which is automatic depression detection from social media posts. This new direction had recently gripped the attention of researchers and developed rapidly (see section II). The researchers were not limited to depression detection, but they extended their studies to detect other mental health disorders like ADHD, Bipolar, Depression, and Anxiety. By time this direction of research had also been extended to include studies of future mental disorders prediction for social media users rather than detecting the already existing disorders [6]. This work, unlike the previous attempts in literature, deploys and compares many Machine learning algorithms (ML) and Deep learning models (DL) to detect and predict mental disorders from REDDIT data. REDDIT is a social news forum or platform that is continuously curated by the site members to cover some important data points that includes but not limited to: post title, post body, comments, time stamps, shares, and scores [7].

In this work the proposed approaches are deployed, compared, and then discussed comprehensively to get the best possible insights and conclusions. REDDIT data was used to validate and evaluate the proposed work. The main advantage of this work over all the state-of-the-art work in literature is the detection of all mental disorders not only depression or anxiety, but it addresses the detection of all kinds of mental disorders which are: ADHD, Bipolar, Anxiety, Depression. The proposed work detects the mental disorder from clinical and non-clinical data, beside that it also predicts the future possibility of mental illness. The second advantage of this work is that it is the first attempt that builds and evaluates more than ten different models to address this problem. The proposed work also used the whole REDDIT data in the period of time starting at 2011 and ending in 2017 without selecting a small sample to avoid any possibility of overfitting.

This work took an extra mile by studying the capabilities of the ML and DL algorithms in the prediction of potential mental disorder not only detecting already existing disorders. The study also covers all mental disorders which are ADHD, Anxiety, Bipolar, not just one or two like literature attempts. This work also represents the first attempt that builds and compares Nineteen different models to produce a solid conclusion about the best models in addressing this kind of research problem.

The rest of this paper is organized as follows: The next section reviews the work done in this area over the last few years. Section III introduces the dataset used in this work and the curation of it. Section IV explains the approach proposed by this paper to detect and predict mental disorders from REDDIT data. Section V demonstrates the experiments done in this work, and discusses the results obtained by the proposed models. Finally, the paper is concluded in section VI, which also suggests future steps to improve results.

II. LITERATURE REVIEW

Recently, the detection of mental disorders from social media posts has grabbed the attention of researchers in the fields of textual data analysis and natural language processing. Some papers focused on the detection of the mental illness, while some other researchers expanded their work to the prediction of future possible illness. Most of the papers focused on a certain type of mental illness, which is depression, while few researchers expanded their work to other types of mental illness like anxiety and mental disorder.

Kumar et al. in 2019 [8], studied the linguistic clues combined with the user posting patterns, i.e., time and frequency of posting on twitter to detect anxious depression from tweets on a real time basis. The authors trained three different classifiers on the sampled tweets of 100 users. The three machine learning classifiers they used to train the data were the Multinomial Naive Bayes, random forest, and Adaboost. Finally, they built an ensemble voter using the three classifiers to classify the 100 users to either anxious depressed or non-anxious depressed, and they reached an accuracy rate of 85%. The main drawback of their work is the small sample of dataset which suggests results overfitting.

In 2019 also, Wongkoblap et al. [9] went to deep learning to automatically detect depression from social media. The authors used Low-Short Term Memory (LSTM) combined with Gated Recurrent Units (GRUs) to perform the task. The authors used 5-fold cross validation, and they reached the maximum accuracy rate at the 2nd and 3rd folds which was 75.49% and an average accuracy rate of 74.65%, those results were relatively not promising enough especially that the authors targeted depression only. In the same year Tariq et al. [10] used the semi supervised learning approach combined with the broadly used supervised ML classifiers like Support Vector Machine (SVM), Naïve Bayes, and Random Forest to detect and classify different mental disorders from social media posts. The authors used Reddit to download posts and their associated comments to train and test their classifiers. The authors confirmed that the combination between the semi supervised approach and the supervised classifiers obtained better results, they also confirmed that the SVM with co-training achieved the better F1-score of 0.84.

But this study was also limited to detecting depression and anxiety based on the detection of negative feelings.

Wongkoblap et al. [9] used the posts on social media also to detect mental health issues. Their study focused on comparing training a predictive model with multiple instance learning (MIL) trained via Long-Short Term Memory (LSTM), with the MIL trained via Convolutional Neural Networks (CNNs). The authors limited their study to the depression symptoms detection, and they confirmed that training an MIL model via LSTMs obtains better accuracy than training the MIL model with CNNs. The authors' study was limited to depression detection, and no specific results were stated.

Rezaii et al. [11] relied on language analysis to detect depression from social media posts. The authors used skip-grams and word2vec to create word embeddings and accordingly they better manipulated large texts. These word embeddings of large texts were fed to a two-layer neural network to analyze text and unpack sentence vectors. The authors mentioned that they achieved a result of 90%, but the authors used a very small sample of dataset which contains only 40 social media participants, the fact that suggests results overfitting. Buddhitha and Inkpen [12] in 2019 also studied the detection of depression and posttraumatic stress disorder (PTSD) from social media posts but this time using deep learning approaches. They used multi class learning with CNNs with multiple channels and multiple inputs like age and gender. The authors also built an emotion classifier that takes tweets as input and obtains the emotion category like sad, fear, and joy as an output. The authors confirmed that they achieved the highest accuracy rate of 88% in classifying emotions using Multi-Channel CNN (MCCNN). But this cannot be considered as a reliable judge of depression diagnosis, as the authors relied on sadness and joy to relate them to depression which cannot indicate an accurate diagnosis of depression.

Thorstad et al. [6] in 2019 also, conducted the first and the only attempt in the state-of-the-art work that addresses all mental disorders which are ADHD, bipolar, anxiety, and depression. The authors used sufficient number of social media posts to evaluate their work including posts from clinical subreddits and non-clinical subreddits. The authors only used logistic regression approach to detect the mental illness and to predict the future occurrence of it, however, they can still be considered the strongest work done in this area, since they didn't only propose the only attempt that addresses the four mental illnesses, but they are also proposing the only attempt that aims at predicting the mental illness before it happens, not only detecting the already existing mental illness. The authors confirmed that they achieved F1 score of 0.74 detecting depression from clinical subreddits, F1 score of 0.44 in detecting the depression from non-clinical subreddits, and F1 core of 0.36 in predicting future possible depression from non-clinical subreddits. Since this work offered the strongest coverage to all coordinates of the problem of mental illness detection and prediction from social media, using the same source of data, we will compare their work to our work at the end of this paper (see section V).

Trifan et al. [13] in 2020, explored the psycholinguistic patterns in social media texts to detect depression. The authors compared three different classifiers combined with Bag of Words (BOW), beside weighting linguistic features using TFIDF. The three classifiers they trained their data on were the Multinomial Naïve Bayes (MNB), Stochastic Gradient Descent (SGD), and Linear Support Vector Machine (SVM). The main contribution added by their study was the Passive Aggressive classifier (PA) that according to their conclusion outperformed the three previously mentioned classifiers. The PA classifier achieved an F1-score of 0.72. In the same year, Jiang et al. [14] studied the linguistic indicators of mental health, and they covered several mental health issues. The authors used BERT to classify among the eight classes, but they achieved average accuracy of about 64% and average F1 score of about 0.645 which cannot be considered good results for such a problem.

Alghamdi et al. [15] collected Arabic texts from social media and used it to study automatic depression detection. The authors compared the Lexicon-based approach using rule-based algorithm, and machine learning based approaches such as Adaboost, K-Nearest Neighbor (KNN), Random Forest (RF), Stochastic Gradient Descent (SGT) and Support Vector Machine (SVM). The authors annotated the data and trained the classifiers with the help of a psychologist to predict depression symptoms. They reported that they exceeded 80% accuracy in depression detection from Arabic posts. Birnbaum et al. [16] used both texts and images on Facebook to detect and identify mental illness. The authors collected 3,404,959 Facebook messages and 142,390 images across 223 participants. The authors evaluated different ways of classification using general purpose classifiers and linear regression, and they reported that they achieved a classification AUC score of 0.77. But using images for such a problem is time and resource consuming meanwhile there is no remarkable leap in results over analyzing text data only.

A year later, Chatterjee et al. [17] used Multinomial Naïve theorem to detect depression from social media posts. The authors reported that their system detected depression with an accuracy rate of only 76.6%. In 2021 also, Ren et al. [18] used the attention model to develop a semantic understanding network to detect depression from Reddit data. The authors relied on understanding and discriminating between negative and positive emotions. They build two units in their network, one for understanding positive emotions and the other one for understanding negative emotions. They limited their study to depression detection, but they reported that their attention network succeeded in detecting depression with an accuracy rate that reached 91.3%. Again, the authors relied on relating negative sentences to depression which cannot be considered an accurate diagnostic technique.

Afterwards, in 2022 Ansari et al. [19] trained many classifiers on text to detect depression. The study mainly focused on the comparison between hybrid and ensemble methods in automatic depression detection from social media posts. The authors confirmed that ensemble methods outperform hybrid classifiers in this area. The authors reported that they got depression detection accuracy of only 75% using ensemble approach and Reddit dataset. Nalini [20] also in 2022 has analyzed the mental health status using Facebook posts and ML classifiers. The author compared K-nearest neighbor (KNN), Support Vector Machine (SVM), and Decision Tree (DT) to classify the post to depressed or not. He reported that DT gives the best results in detecting depression over other previously mentioned classifiers. But at the end, the study was limited to figuring out the challenges related to the problem, and some recommendations to overcome them with no specific results of detection or prediction.

In 2023 Tufail et al. [21] proposed a depression detection approach using convolution neural networks (CNNs), and they confirmed that they achieved a validation accuracy rate of only 64%. The authors then confirmed that they were able to increase the accuracy from 64% to 68% when they used complex data generation and augmentation methods, but this still very low rate of accuracy compared to other work done in literature. In the same year, Koushik et al. [22] built and tested three different models to detect the signs of depression from social data. The first model was the SVM, the second one was the CNNs, and the third model was the BI-LSTM. SMOTE was used in the three models for dataset oversampling. The authors confirmed that the SVM was the champion model as it outperformed the results of the two other models with an accuracy rate that reached only 60% which is very low.

Hasib et al. [23] in 2023 have surveyed the state-of-the-art Machine learning and deep learning approaches used to detect depression from social media and they confirmed that ML and DL can share in efficient diagnosis of depression using personal status posted on social media. Yicheng et al. [24] collected their data from the Chinese social network platform Sina Weibo. The authors proposed a feature section method for analyzing depression symptoms using Multivariate time series approach, but the authors ended up with only correlating the disease to some of its symptoms. Li et al. [25] built and deployed a multimodal attention mechanism for classifying social media users to depressed or normal users. The authors confirmed that analyzing the text data with picture added to it can lead to better results in depression detection, but this is not always the case on social media, beside some other drawbacks of their prosed solution such as the complexity and time consumed performing classification task.

In 2024, Helmy et al. [26] extended their classification to anxiety, depression, and normal social media users. The authors also performed the task on both English and Arabic texts, and they confirmed achieving good accuracy results, but their dataset was very small as they used only 10,000 Arabic tweets and about 60,000 English tweets. The most recent attempt done in this area was the attempt done by Dhariwal et al. [27], where the authors used and compared several machine learning, ensemble learning, and deep learning algorithms to automatically detect the mental disorder from social media data. The authors confirmed that CNN outperformed all the traditional machine learning algorithms and ensemble learning methods with an accuracy rate that reached 99.7%. But the attempt was just a pilot study that used 'Cities Health Initiative Dataset', not real posts for real users, to study the capabilities of machine learning and deep learning models in that area.

III. DATASETS

This section demonstrates the nature and characteristics of the different datasets used in this work. Three different studies have been conducted in this work, in which the main goal of first study is to determine if machine learning algorithms (ML), ensemble learning algorithms (EL), and large language models (LLMs) could detect if a person suffers from a mental illness disorder from his posts in a clinical context in social media. The main goal of second study is to detect if a person suffers from a mental illness disorder from his posts but this time in a nonclinical context on social media using the same algorithms. The main goal of third study is to determine if machine learning algorithms could be used to predict the future occurrence of mental illness before a person has enough awareness of his/her case, and this is the most novel part in this work as this is the second attempt aims at predicting mental disorder before it happens.

In the three studies, all mental disorders which are ADHD, Anxiety, Bipolar, and Depression were considered, and this is another added value for this work, as this is the second study that aimed at detecting and predicting all mental disorders not just one or two of them. This work also represents the first attempt that builds and compares wide diversity of models to detect and predict mental disorder in different contexts of social media data (see section IV).

REDDIT social media platform was chosen to train and evaluate the proposed approaches in this work to guarantee the largest possible pool of social media data, and accordingly guarantee avoiding results overfitting. In all studies the dataset was divided into training and testing datasets with the ratio of 80% to 20% respectively.

A. DATA ACQUISTION FROM CLINICAL SUBREDDITS

The REDDIT application programming interface was used to download and collect the data used to conduct this study. The posts downloaded represent Seven years of social media posts from 2011 to 2017. The posts were then randomly under sampled to create a balanced dataset of 41,861 posts for each disorder. Stratified random sampling [28] were used to have same percentage for each disorder in training and testing sets. In the training set there are 33,489 posts for each disorder and in the testing set there are 8,372 posts for each disorder,

i.e. the total number of posts used in this study was 167,444 clinical social media posts.

B. DATA ACQUISTION FROM NON-CLINICAL SUBREDDITS

The aim of the second study is to see whether the proposed models can detect mental disorder from non-clinical data as mentioned before. Here users' posts were downloaded using REDDIT application programming interface also, where for each user, all the user's posts were concatenated into a single data point to avoid the same user appearing in the training and testing dataset (which could artificially inflate accuracy rates based the same user having a consistent but idiosyncratic linguistic style).

3,252,035 user posts for 20,914 users were downloaded, then a random data under-sampling was held to have a balanced dataset with 19,276 users (4,819 user per disorder). Stratified random sampling was used to have the same percentage for each user in training and testing sets. In the training set there are 3,855 users and, in the testing set there are 964 users. At the end, the total number of posts used in this study was 2,987,780 non-clinical social media posts, which is the largest number of posts used for this task compared to all work in literature.

For the future prediction of mental disorder study, the number of posts downloaded were expanded to include users' posts in non-clinical subreddits before he/she ever posted in the clinical subreddits. 660,844 users' posts were downloaded for 15,100 users. Afterwards a random under-sampling was held again to have balanced dataset among all disorders data in which a total number of 12,572 users were distributed to 3,143 users per disorder. Stratified random sampling was then used to have the same percentage for each user in training and testing sets. In the training set there are 2,514 users' posts and, in the testing set there are 629 users posts per disorder. At the end, the total number of posts used in this study was 1,948,660 non-clinical social media posts.

C. DATA PREPROCESSING

Regarding the preprocessing steps applied to the raw posts to prepare it to be fed to the proposed models. Firstly, noise removal and text normalization were held to standardize the format of the text to enhance the model's performance and generalization capability. Afterwards, leading and trailing whitespace is meticulously removed to prevent any inadvertent interference with subsequent processing steps. Explicit mentions of the disorders were then removed from the text and their prefixes by removing the words beginning with anxiety, depression, bipolar, and ADHD. Then any non-alphabetic characters, such as punctuation marks and special symbols from the text were removed. Then, a final stripping operation is judiciously executed to eradicate any residual whitespace. Finally, each post was converted to Tf-idf representation in case of ML and Ensemble learning studies, meanwhile, the posts were converted to embedding vectors in case of Language Models study. Figure 1 shows the steps of data collection and preparation in detail.

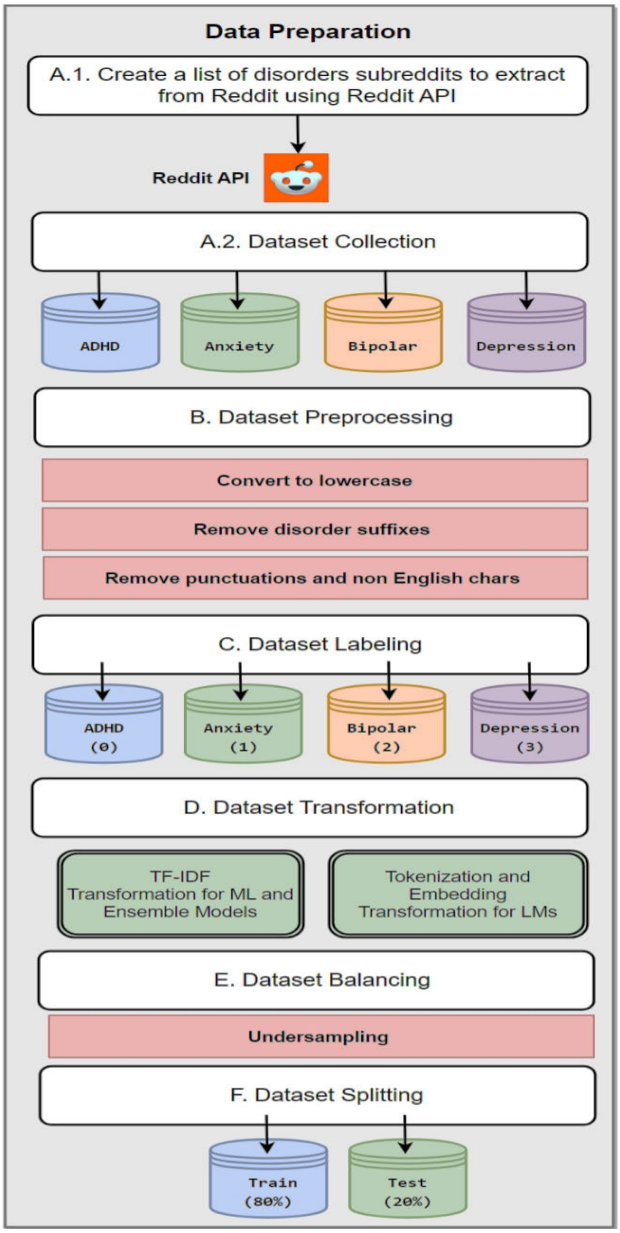

FIGURE 1. The steps of data collection and preparation.

IV. THE PROPOSED APPROACH

This is the first attempt in literature that aims at both detecting all types of disorders and predicting them before they appear using this large number of algorithms. In this work Six different machine learning classifiers, Nine ensemble learning classifiers, and four different LLMs were deployed to address the automatic detection and future prediction of mental disorder from social media posts in both clinical and non-clinical subreddits. Figure 2 summarizes the general steps used in all machine learning and ensemble learning models deployed and tested in this work. Figures 3 and 4 summarize the settings and the training steps of the language models.

Machine Learning classifiers deployed, trained, and tested in this work were Logistic regression [29], Support Vector

Algorithm:

1. **Collect Data** from Reddit (Clinical & Non-clinical).
2. **Remove** words related to the four mental disorders.
3. **Under-sample the data** to balance it among the four classes.
4. **Split** data to 80% for training and 20% for testing.
5. **Set the hyperparameters** to the model's most default values.
6. **Train** the model.
7. **Test** the model capability of existing mental disorder classification, and potential mental disorder prediction.
8. **Compute the F1-score** for every mental disorder.
9. **Compute the Average F1-score** to evaluate the model general performance for all mental disorders.
10. **Compare** the results obtained by the model against other models used in this work.
11. **Compare** the champion models with the state-of-the-art models in literature.

FIGURE 2. The general steps of the ML and EL models used in this work.

Encoder only Models Architecture:

1. **Use a multi-layer** bidirectional Transformer **encoder.**
2. **Pre-train the model** using unsupervised tasks:
   1. Masked Language Model (MLM): Randomly mask some tokens and predict them based on context.
   2. Next Sentence Prediction (NSP): Predict if the second sentence follows the first in a pair (only for BERT).

Encoder only Models Processing & Training

1. **Tokenize input text** into subwords
2. **Add special tokens** to mark the beginning and end of the sentence.
3. **Pass tokenized input** through the model.
4. **Pass the output** of the classification tokens to:
   1. A classification layer that contains four nodes (one for each class).
   2. A Softmax activation function.

FIGURE 3. The general settings and training steps of the encoder only language models.

Machine (SVM) [30], K-nearest neighbors (KNN) [31], Decision Tree [32], Stochastic Gradient Descent (SGD) [33], and Multinomial Naive Bayes [34]. In the ensemble learning category, several models were applied to do the same task such as Voting Classifiers [35], Random Forest [36], Bagging Meta-Estimator [37], AdaBoost [38],

Decoder only Models Architecture:

1. **Use decoder-only** Transformer.
2. **Pre-train the model** using an autoregressive language modeling objective (predicting the next word in a sequence).

Decoder only Models Processing & Training

1. **Tokenize input text** into subwords.
2. **Pass tokenized input** through the model.
3. **Compute the mean** of the output of the last hidden state and pass it to:
   1. A classification layer that contains four nodes (one for each class).
   2. A Softmax activation function.

FIGURE 4. The general settings and training steps of the decoder only language models.

XGBoost [39], Gradient Boosting [40], and Light Gradient Boosting Machine (LightGBM) [41]. In the deep learning category, different large language models (LLMs) were used such as Bert [42] & [43], RoBERTa [44] & [45], GPT [46] & [47], and GPT2 [48]. The following figures demonstrate the architecture of the prosed models. Figure 5 summarizes the set of models used in this work. Figure 6 shows the architecture of the first heterogenous voting classifier (VC1), figure 7 represents the architecture of the second voting classifier (VC2), and figure 8 shows the architecture of the third one (VC3). Figure 9 shows the BERT architecture which is similar to the RoBERTa architecture as the difference is just in the amount of data and the training time as RoBERTa requires more data and more training time. Figure 10 shows the GPT architecture which is similar to the architecture of the Open AI GPT, where again the only difference is in the size of the dataset fed into the model and the training time, as GPT2 requires more data and more training time compared to OpenAI GPT.

The following subsections demonstrate parameters settings for each model used in detecting and predicting mental disorders from clinical data and non-clinical data. Most of the hyperparameters used in this work were set to the default values mainly for two reasons. The first reason was the scalability and generalization purposes. The second one was the constraints of time, memory, and computational resources.

Scikit-learn library was used to implement the ML classifiers and the ensemble learning models mentioned before. TensorFlow was used to train the large language models (LMMs) used to handle the same task.

1) CLASSICAL ML SETTINGS

This study started by training a logistic regression (LR) model by considering the huber penalty, C, multi class, max iter, and

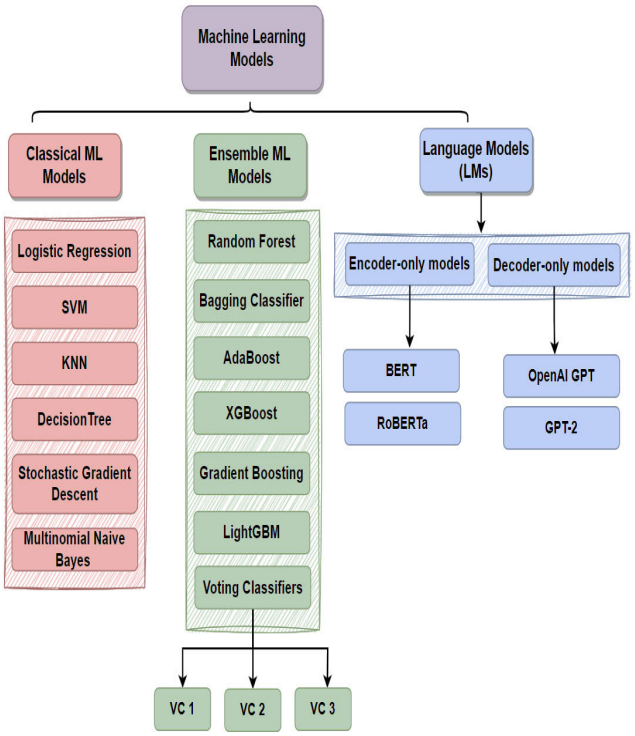

FIGURE 5. Summary of models used in mental disorder detection and prediction.

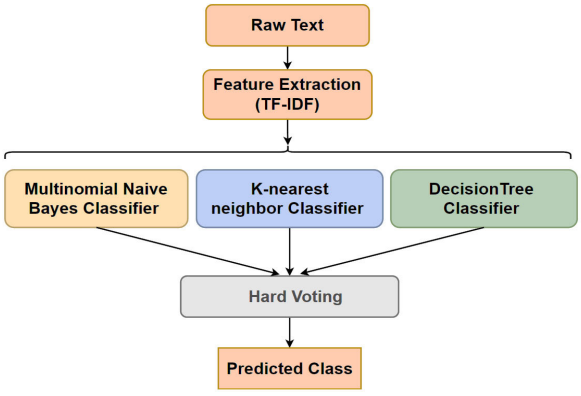

FIGURE 6. The architecture of the first voting classifier (VC1).

random state with values L2, 1.0, One Versus Rest (OVR), 1000, and 0 respectively. The key factor in choosing L2 instead of L1 is its applicability to handle a dataset of high dimensionality in an efficient way without the need of feature selection and reduction. L1 in case of handling such a huge dataset (millions of social posts) would need feature selection which consumes a lot of time and computational power which is not available at this stage of work to avoid overfitting.

The support vector machine (SVM) with random state set to 0. The K-nearest neighbors (KNN) model with n neighbors set to 5. The random state of the decision tree is set to 0. In the stochastic Gradient Descent (SGD) model, the parameters were set as follows: hyper-parameters loss, penalty, and random state with these values modified huber, l2, and

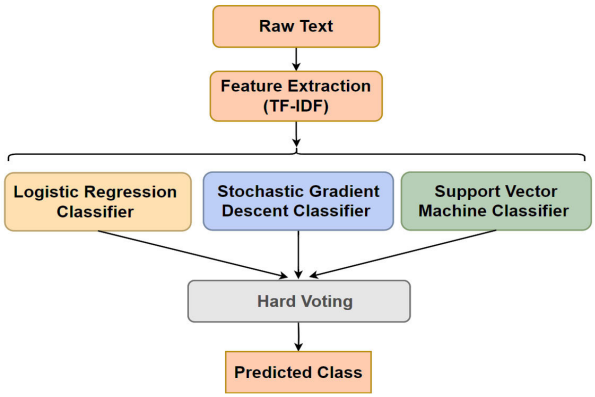

FIGURE 7. The architecture of the second voting classifier (VC2).

0 respectively. The alpha of the Multinomial Naive Bayes (MNB) model is set to 0.001.

2) ENSEMBLE LEARNING SETTINGS

Three heterogenous voting classifier models were built in this study by combining different traditional ML classifiers, the first voting classifier (VC1) was built using k-nearest neighbors (KNN), decision tree (DT), and Multinomial Naive Bayes (MNB) models, the second one (VC2) using logistic regression, SVM, and SGD, and the third one (VC3) using logistic regression (LR), random forest (RF), SVM, DT, MNB, and SGD.

In the first two voting classifier models, hard voting strategy [49] was used, meanwhile in the third voting classifier model, soft voting strategy was used.

For the first voting classifier model (VC1), three traditional classifiers were stacked together; k-nearest neighbors (KNN) with n neighbors set to 5, the decision tree with a random state set to 0, and Multinomial Naive Bayes with an alpha set to 0.001. In the second voting classifier (VC2) three different traditional classifiers were stacked together; logistic regression (LR) by considering the huber loss, penalty, C, multi class, max iter, and random state with these values l2, 1.0, One Versus Rest (OVR), 1000, and 0 respectively, SVM with a random state set to 0, and SGD used the following huber loss, penalty, and random state with these values modified huber, l2, and 0 respectively. In the third voting classifier (VC3), all classifiers involved in the first two voting classifiers were stacked together, with the same hyper-parameters, but this time soft voting strategy was deployed.

Besides the three voting classifiers mentioned before, several bagging and boosting well-known ensemble models were trained and tested to find out which model would be the best one for this task. The settings used in this work for those ensemble models are as follows: The random forest model with a random state set to 0. The bagging meta-estimator model with a base estimator set to logistic regression with the following hyperparameter penalty, C, multi class, max iter, and random state with these values l2, 1.0, One Versus Rest (OVR), 1000, and 0 respectively, N estimators set to 1000, and random state equal to 0. The AdaBoost model, with a

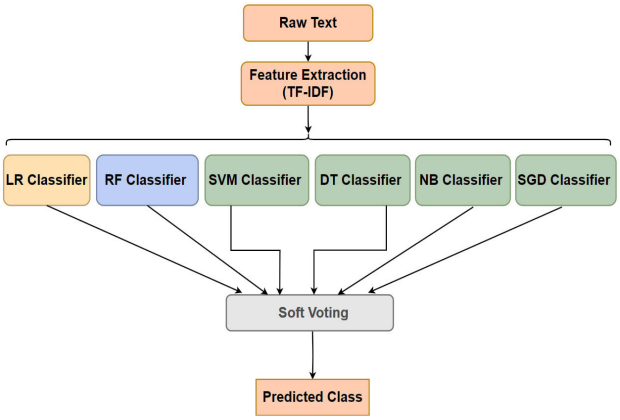

FIGURE 8. The architecture of the third voting classifier (VC3).

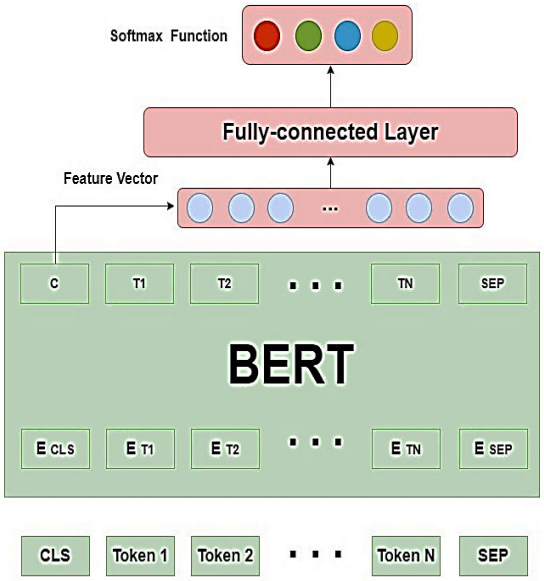

FIGURE 9. BERT architecture.

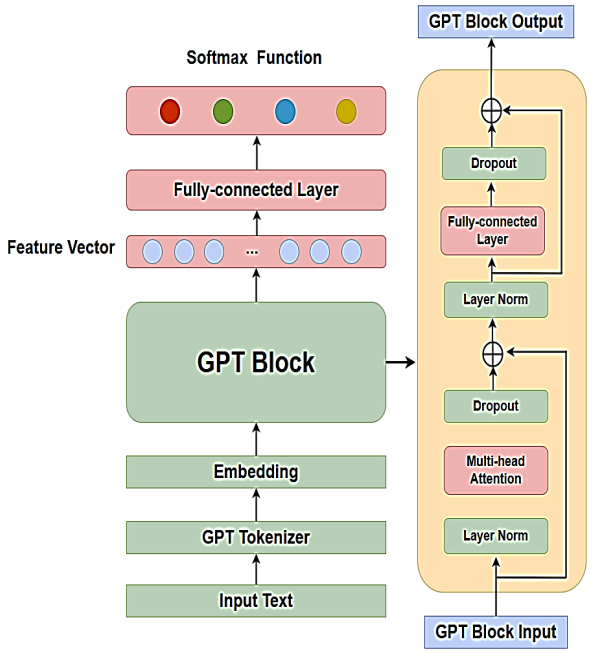

FIGURE 10. GPT architecture.

random state set to 0, base estimator set to Stochastic Gradient Decent (SGD) with the following hyper-parameters loss, penalty, and random state with these values modified huber, l2, and 0 respectively, algorithm set to SAMME, learning rate set to 0.09, and n estimators set to 500. The random state of the XGBoost is set to 0. The learning rate of Gradient Boosting model is set to 0.01, and its random state is set to 0. The n estimators and the random state of the LightGBM model were set to 500 and 0 respectively.

3) LANGUAGE MODELS SETTINGS

In this important part of our study, four large language models with two different architectures were trained and tested for the same tasks. Encoder only models and decoder-only models. Bert and Roberta were selected for Encoder-only models and OpenAI GPT and GPT 2 were selected for Decoder-only models. In this part, the Hugging-face ecosystem was used to implement the language models mentioned before.

Each language model was trained using TensorFlow framework. Adam optimizer was used with a learning rate equal to 2e-5. For the BERT model, the default hugging-face model configuration was used with max position embedding, the number of attention layers, and the number of attention heads equals 512, 12, and 12 respectively. For the RoBERTa model, the default hugging face model configuration was used with max position embedding, the number of attention layers, and the number of attention heads equals 514, 12, and 12 respectively. For the OpenAI GPT model, the default hugging face model configuration was used with max position embedding, the number of attention layers, and the number of attention heads equals 768, 12, and 12 respectively. For the GPT2 model, the default hugging-face model configuration was used with max position embedding, the number of attention layers, and the number of attention heads equals 768, 12, and 12 respectively.

The hardware Specifications used in this work were as follows: For Classical and ensemble ML models: 12th Gen Intel(R) Core (TM) i7-12650H processor, 2.30GHz 32 GB System Ram, NVIDIA GPU, GeForce RTX 3060, 6 GB GPU Ram, and Hard disk size of 1 TB. For LLMs: Intel(R) Xeon(R) CPU, 2.20GHz, 83.5 GB System Ram, A100 GPU, 40 GB GPU Ram, and Hard disk size of 201 GB.

The longest training time was taken by the language models which reached an average training time of 10 hours.

V. RESULTS AND DISCUSSION

This section discusses the experiments held in this work to evaluate the models deployed to detect and predict metal disorders. The following subsections discusses the results of the proposed models per each study of the three studies conducted in this work which are: 1) Detection of mental disorder from clinical data, 2) detection of mental disorder

from non-clinical data, and 3) future prediction of potential mental disorder from non-clinical data.

A. FIRST STUDY: CLINICAL DOMAIN SUBREDDITS

A wide range of experiments have been conducted in this study. Those experiments were mainly divided into three main categories: classical machine learning experiments, ensemble learning experiments, and deep learning experiments (see section IV).

Because a classifier can be accurate while failing one of the important objectives, which are the precision (PR) and recall (R) the model obtains—for example, by learning to guess the class with the highest base rate— although the data was sampled to have balanced classes, the F1 score is typically selected over accuracy, as the F1 score combines both precision and recall in one metric. The following equation shows how the F1 score is calculated.

$$F1\ score = 2 \times (PR \times R)/(PR + R) \quad (1)$$

where:

PR ≡ Precision of the model

R ≡ Precision of the model

To calculate the average F1 score over all classes for a certain model. The following equation shows how the average F1 score for every model was calculated.

$$Avg\ F1 = (F1_{C1} + F1_{C2} + F1_{C3} + F1_{C4})/4 \quad (2)$$

where:

Avg F1 ≡ Average F1 score of the model overall classes.

$F1_{C1}$ ≡ F1 score of the model for class 1.

Here every class represents a mental disorder, i.e. depression, anxiety, etc.

1) CLASSICAL ML RESULTS

Logistic regression (LR) set to the parameters mentioned in the previous section succeeded in achieving an f1 score of 0.82, 0.74, 0.74, and 0.73 in the detection of ADHD, Anxiety, Bipolar, and Depression disorders respectively, with an average F1-score of 0.76. The support vector machine (SVM) with random state set to 0 as mentioned before, achieved an f1-score of 0.82, 0.74, 0.74, and 0.73 in the detection of ADHD, Anxiety, Bipolar, and Depression disorders respectively, with an average F1-score of 0.76.

The K-nearest neighbors (KNN) model with N equal to 5 achieved an F1-score of 0.07, 0.41, 0.04, and 0.15 in the detection of ADHD, Anxiety, Bipolar, and Depression disorders respectively, with an average F1-score of 0.17 which is the worst amongst all used classifiers (see fig. 4).

The decision tree with a random state equal to 0 achieved an F1-score of 0.62, 0.52, 0.57, and 0.52 in the detection of ADHD, Anxiety, Bipolar, and Depression disorders respectively with an average F1-score of 0.56.

The stochastic Gradient Descent (SGD) model with the hyper parameters set as mentioned before achieved an f1-score of 0.82, 0.74, 0.74, and 0.73 in the detection of ADHD, Anxiety, Bipolar, and Depression disorders respectively with an average F1-score of 0.76. The Multinomial Naive Bayes (MNB) model with an alpha set to 0.001 achieved an F1-score of 0.76, 0.65, 0.66, and 0.67 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively, with an average F1-score of 0.69. Logistic regression (LR), SVM, and SGD succeeded in obtaining the highest results in detecting mental disorders from clinical data with an equal average F1-score of 0.76. Figure 11 compares the average F1-scores obtained by the ML classifiers used in the first study.

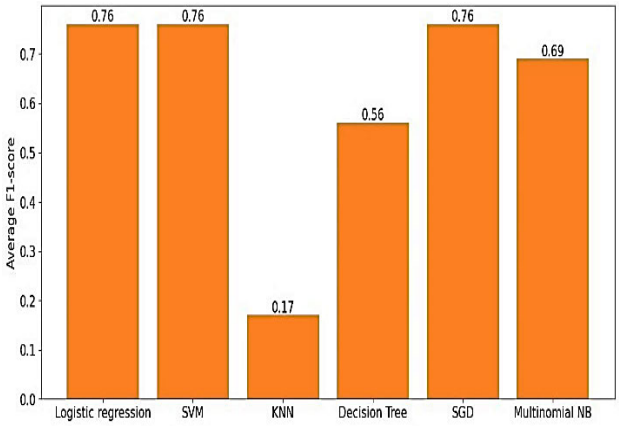

FIGURE 11. Study 1: F1-score for classical machine learning classifiers.

2) ENSEMBLE LEARNING RESULTS

The first voting classifier VC1 (KNN + DT + MNB) with hard voting strategy, achieved an F1-score of 0.69, 0.58, 0.57, and 0.55 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively with an average F1-score of 0.60. The second voting classifier model VC2 (LR + SVM + SGD) with hard voting strategy, achieved an F1-score of 0.82, 0.74, 0.74, and 0.73 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively, with an average F1-score of 0.76. The third voting classifier VC3 (LR + RF +SVM +DT + MNB + SGD) with soft voting strategy, achieved an F1-score of 0.81, 0.72, 0.73, and 0.71 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively, with an average F1- score of 0.75.

The random forest model achieved an F1-score of 0.77, 0.67, 0.68, and 0.68 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively with an average F1-score of 0.70. The Bagging Meta Estimator (BME) achieved an F1-score of 0.82, 0.74, 0.74, and 0.73 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively, with an average F1-score of 0.76.

The AdaBoost model achieved an f1-score of 0.78, 0.71, 0.70, and 0.70 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively, with an average F1-score of 0.72. The XGBoost model achieved an F1-score of 0.78, 0.71, 0.72, and 0.70 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively, with an average F1-score of 0.73. The Gradient Boosting model (GB) achieved an F1-score of 0.55, 0.52, 0.53, and 0.57 in the detection of

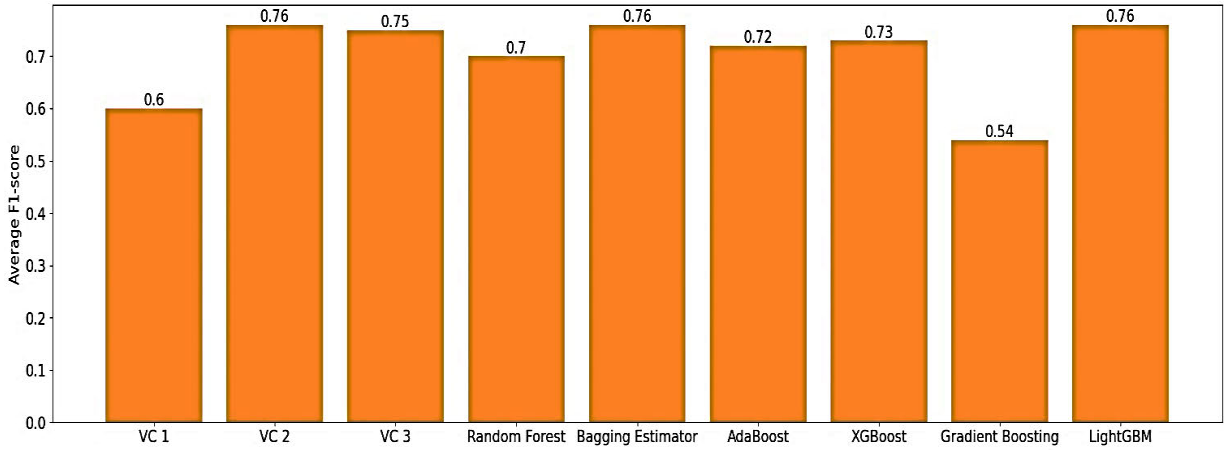

FIGURE 12. Study 1: F1-score for ensemble learning models.

ADHD, Anxiety, Bipolar, and Depression respectively with an average F1-score of 0.54. The LightGBM model achieved an F1-score of 0.82, 0.74, 0.75, and 0.72 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively with an average F1-score of 0.76. From the above results, VC2, Bagging Meta Estimator (BME), and LightGBM proved to be the best amongst the rest used ensemble models in detecting different mental disorders from clinical data where the three models achieved an average F1 score of 0.76 in this task. Figure 12 summarizes the Micro F1 scores obtained by all ensemble models in the first study.

3) LANGUAGE MODELS RESULTS

BERT achieved an F1-score of 0.86, 0.79, 0.79, and 0.76 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively with an average F1-score of 0.80. RoBERTa model achieved an F1-score of 0.87, 0.79, 0.78, and 0.76 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively with an average F1-score of 0.80.

OpenAI GPT model achieved an F1-score of 0.86, 0.78, 0.79, and 0.76 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively with an average F1-score of 0.80. GPT2 model achieved an F1-score of 0.86, 0.78, 0.77, and 0.76 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively with an average F1-score of 0.79. Figure 13 compares the average F1-scores obtained by the LLMs used in the first study.

As can be shown in the previous discussion Large Language Models (LLMs) outperformed the classical machine learning classifiers and the ensemble learning models in the task of detecting mental disorders from clinical text data or subreddits. The champion models overall the LLMs used in this study are BERT, RoBERTa, and OpenAI GPT. The champion models were chosen based on the average F1-score taken from the F1-scores obtained for every disorder, but some models were better than others per every disorder. BERT proved to be the best in detecting anxiety and bipolar, meanwhile RoBERTA proved to be the best in detecting ADHD, and OpenAI GPT proved to be the best in detecting depression (see table 1).

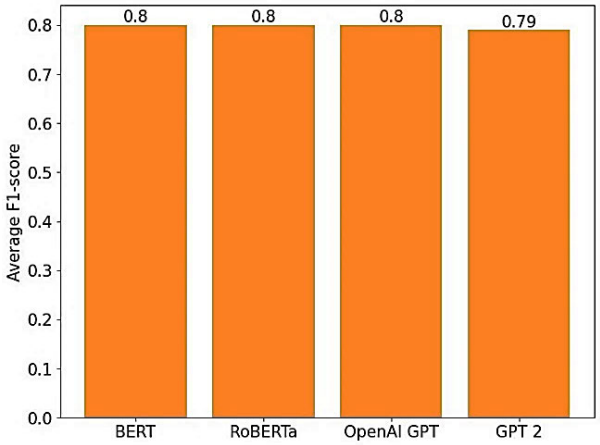

FIGURE 13. Study 1: F1-score for large language models.

B. SECOND STUDY: NON-CLINICAL DOMAIN SUBREDDITS

In this study, the experiments that were conducted in the first study are repeated, but this time for detecting mental disorders from non-clinical social media posts.

1) CLASSICAL ML RESULTS

The same machine-learning algorithms with the same hyper-parameters used in the first study are used here in this study. The experiments in this part started by training and testing the logistic regression model. It succeeded in achieving an F1 score of 0.52, 0.47, 0.51, and 0.57 in the detection of ADHD, Anxiety, Bipolar, and Depression disorders respectively, with an average F1-score of 0.52. The support vector machine (SVM) achieved an F1-score of 0.50, 0.46, 0.51, and 0.55 in the detection of ADHD, Anxiety, Bipolar, and Depression disorders respectively, with an average F1-score of 0.51.

The K-nearest neighbors (KNN) achieved an F1-score of 0.34, 0.30, 0.19, and 0.40 in the detection of ADHD,

TABLE 1. Summary of the results obtained by the models in every category of approaches (machine learning classifiers, ensemble learning models, and large language models) in every study, with champion models results highlighted using bold font.

Study #	Category	Champion Model	ADHD	Anxiety	Bipolar	Depression	AVG F1 Score
#1 Detection from Clinical subreddits	Machine Learning	LR	0.82	0.74	0.74	0.73	0.76
		SVM	0.82	0.74	0.74	0.73	
		SGD	0.82	0.74	0.74	0.73	
	Ensemble Learning	VC2	0.82	0.74	0.74	0.73	0.76
		Bagging Est.	0.82	0.74	0.74	0.73	
		LightGBM	0.82	0.74	0.75	0.72	
	Large Language Models	BERT	0.86	**0.79**	**0.79**	**0.76**	**0.80**
		RoBERTa	**0.87**	**0.79**	**0.79**	**0.76**	
		Open AI GPT	0.86	0.78	**0.79**	**0.76**	
#2 Detection from non-clinical subreddits	Machine Learning	LR	**0.52**	0.47	0.51	**0.57**	**0.52**
	Ensemble Learning	Bagging Est.	0.52	0.47	0.51	**0.57**	**0.52**
		XGBoost	**0.52**	0.47	**0.54**	0.53	
		LightGBM	0.50	**0.49**	0.53	0.55	
	Large Language Models	Open AI GPT	0.43	0.30	0.34	0.50	0.45
#3 Prediction from non-clinical subreddits	Machine Learning	LR	0.41	**0.35**	**0.43**	**0.51**	**0.43**
	Ensemble Learning	Bagging Est.	**0.42**	**0.35**	**0.43**	0.50	**0.43**
	Large Language Models	RoBERTa	0.45	0.32	0.44	0.45	0.42

Anxiety, Bipolar, and Depression disorders respectively, with an average F1-score of 0.31. The decision tree (DT) achieved an F1-score of 0.35, 0.34, 0.37, and 0.36 in the detection of ADHD, Anxiety, Bipolar, and Depression disorders respectively with an average F1-score of 0.36. The stochastic Gradient Descent (SGD) achieved an F1- score of 0.49, 0.46, 0.49, and 0.52 in the detection of ADHD, Anxiety, Bipolar, and Depression disorders respectively with an average F1-score of 0.49. The Multinomial Naive Bayes model achieved an F1- score of 0.45, 0.41, 0.43, and 0.53 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively, with an average F1-score of 0.46. Logistic regression proved to be the best amongst the six machine learning classifiers in detecting and classifying mental disorders using non-clinical social media data. Figure 14 compares the average F1-scores obtained by the ML classifiers used in the second study.

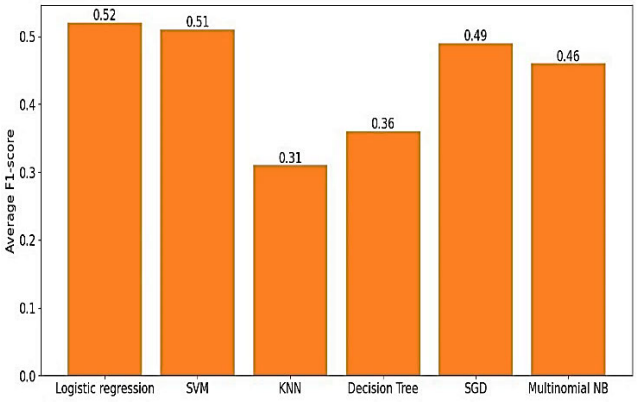

FIGURE 14. Study 2: F1-score for classical machine learning classifiers.

2) ENSEMBLE LEARNING RESULTS

Again, the same ensemble-learning algorithms with the same hyperparameters used in the first study are used here. The experiments in this part started by training and testing the three prementioned voting classifiers VC1, VC2, and VC3. VC1 achieved an F1-score of 0.44, 0.36, 0.32, and 0.46 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively with an average F1-score of 0.40. VC2 achieved an F1-score of 0.51, 0.46, 0.51, and 0.56 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively, with an average F1- score of 0.51. VC3 achieved an F1-score of 0.48, 0.44, 0.48, and 0.52 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively, with an average F1-score of 0.48. The random forest model achieved an F1-score of 0.42, 0.37, 0.42, and 0.49 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively with an average F1-score of 0.43.

The bagging meta-estimator model achieved an F1-score of 0.52, 0.47, 0.51, and 0.57 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively, with an average F1-score of 0.52. The AdaBoost model achieved an f1-score of 0.51, 0.47, 0.50, and 0.55 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively, with an average F1-score of 0.51. The XGBoost model achieved an F1-score of 0.52, 0.47, 0.54, and 0.53 in the detection of ADHD,

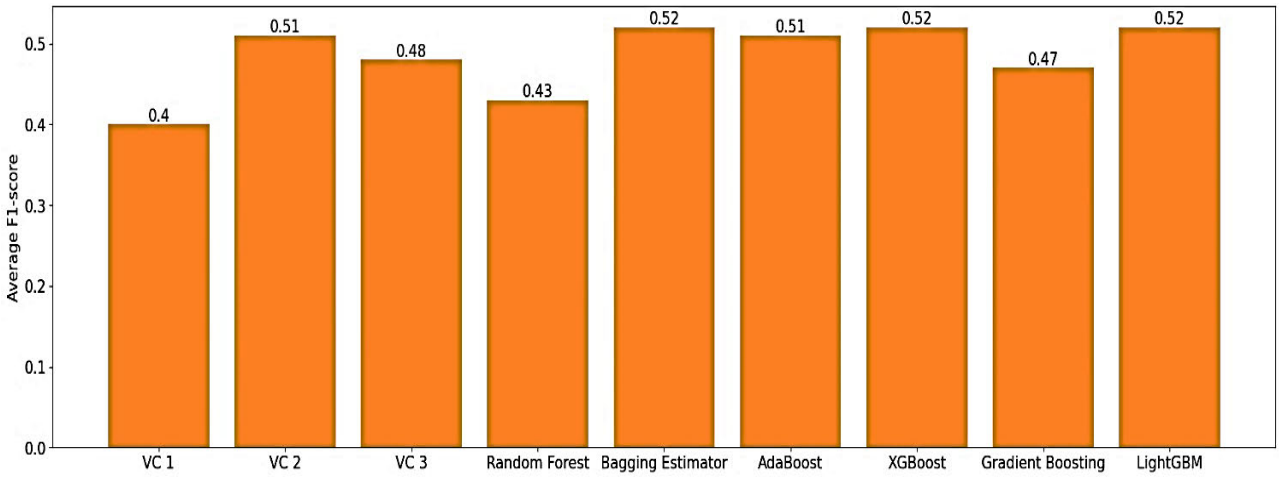

FIGURE 15. Study 2: F1-score for ensemble learning models.

Anxiety, Bipolar, and Depression respectively, with an average F1-score of 0.52. The Gradient Boosting model achieved an F1-score of 0.45, 0.43, 0.49, and 0.49 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively with an average F1-score of 0.47. The LightGBM model achieved an F1-score of 0.50, 0.49, 0.53, and 0.55 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively with an average F1-score of 0.52. Bagging estimator and XGBoost outperformed the other used ensemble learning models used in the detection and classification of mental disorders from non-clinical social media data. Figure 15 summarizes the average F1 scores obtained by all ensemble models in the second study.

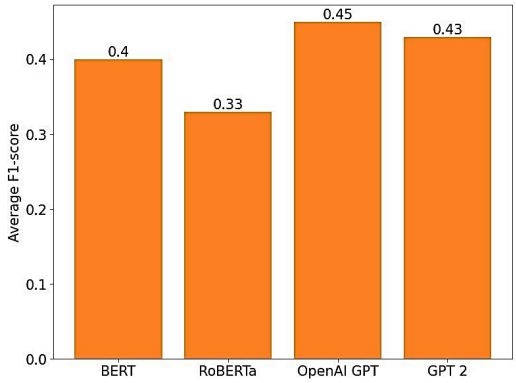

FIGURE 16. Study 2: F1-score for large language models.

3) LANGUAGE MODELS RESULTS

The same LLMs with the same hyperparameters used in the first study were used here also. The BERT model achieved an F1-score of 0.46, 0.37, 0.25, and 0.51 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively with an average F1-score of 0.40. The Roberta model achieved an F1-score of 0.44, 0.35, 0.04, and 0.50 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively with an average F1-score of 0.33.

OpenAI GPT model achieved an F1-score of 0.46, 0.40, 0.43, and 0.50 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively with an average F1-score of 0.45. The GPT2 model achieved an F1-score of 0.36, 0.42, 0.46, and 0.46 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively with an average F1-score of 0.43. OpenAI GPT proved to be the best amongst the other four LLMs in the task of detection and classification of mental disorders from non-clinical social media posts. Figure 16 compares the average F1-scores obtained by the LLMs used in the second study.

Surprisingly ML classifiers and ensemble learning models have outperformed Large Language Models (LLMs) in the task of detecting mental disorders from non-clinical text data or subreddits. The champion model overall the ML classifiers is the logistic regression (LR), and the champion models overall the ensemble learning models used in this study are Bagging estimator, XGBoost, and LightGBM. Again, the champion models were chosen based on the average F1-score calculated from the F1-scores obtained for all disorders. The champion models for every disorder were as follows: LR and Bagging estimator are the best in detecting depression, XGBoost is the best in detecting anxiety and bipolar, and LightGBM is the best in detecting ADHD.

C. THIRD STUDY: FUTURE DISORDER PREDECTION

The same experiments with the same models were conducted here, but this time to predict the future possible mental disorder from non-clinical social media posts before it happens.

1) CLASSICAL ML RESULTS

The same machine-learning algorithms with the same settings of hyperparameters were examined in this study also. The experiments in this part started by training a logistic regression model. It succeeded in achieving an f1 score of 0.41, 0.35, 0.43, and 0.51 in the detection of ADHD, Anxiety, Bipolar, and Depression disorders respectively, with an average F1-score of 0.43. The support vector machine (SVM)

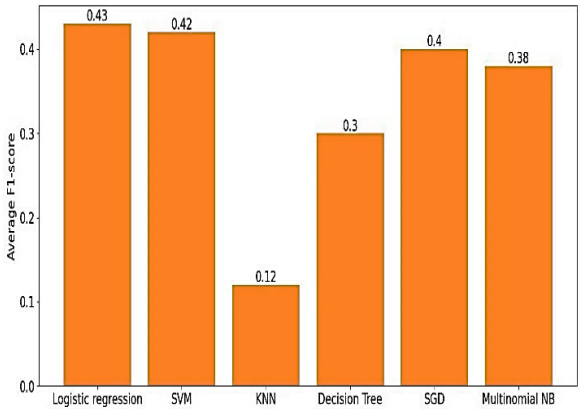

FIGURE 17. Study 3: F1-score for classical machine learning classifiers.

achieved an f1-score of 0.43, 0.33, 0.43, and 0.49 in the detection of ADHD, Anxiety, Bipolar, and Depression disorders respectively, with an average F1-score of 0.42.The K-nearest neighbors (KNN) achieved an F1-score of 0.03, 0.02, 0.41, and 0.03 in the detection of ADHD, Anxiety, Bipolar, and Depression disorders respectively, with an average F1-score of 0.12. The decision tree achieved an F1-score of 0.30, 0.25, 0.31, and 0.32 in the detection of ADHD, Anxiety, Bipolar, and Depression disorders respectively with an average F1-score of 0.30. The stochastic Gradient Descent (SGD) achieved an f1-score of 0.35, 0.37, 0.42, and 0.46 in the detection of ADHD, Anxiety, Bipolar, and Depression disorders respectively with an average F1-score of 0.40.

The Multinomial Naive Bayes model achieved an F1-score of 0.39, 0.32, 0.37, and 0.44 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively, with an average F1-score of 0.38. Figure 13 compares the average F1-scores obtained by the ML classifiers used in the third study. Logistic regression has proved to be the best amongst all used machine learning classifiers in the prediction of potential future mental disorder. Figure 17 compares the average F1 score of the six machine learning classifiers.

2) ENSEMBLE LEARNING RESULTS

The same ensemble-learning algorithms with the same hyperparameters were used here also. The experiments in this part started by training three voting classifier models VC1, VC2, and VC3 discussed before. VC1 achieved an F1-score of 0.37, 0.26, 0.41, and 0.29 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively with an average F1-score of 0.33. VC2 achieved an F1-score of 0.41, 0.34, 0.42, and 0.50 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively, with an average F1- score of 0.42. VC3 achieved an F1-score of 0.39, 0.31, 0.39, and 0.45 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively, with an average F1- score of 0.39. The random forest model achieved an F1-score of 0.39, 0.27, 0.40, and 0.42 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively with an average F1-score of 0.37. The bagging meta-estimator model achieved an F1-score of 0.42, 0.35, 0.43, and 0.50 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively, with an average F1-score of 0.43. The AdaBoost model achieved an f1-score of 0.41, 0.35, 0.41, and 0.48 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively, with an average F1-score of 0.41. The XGBoost model achieved an F1-score of 0.40, 0.32, 0.41, and 0.47 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively, with an average F1-score of 0.40. The Gradient Boosting model achieved an F1-score of 0.41, 0.25, 0.40, and 0.43 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively with an average F1-score of 0.37. The LightGBM model achieved an F1-score of 0.40, 0.35, 0.41, and 0.45 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively with an average F1-score of 0.40. In the study of prediction of mental disorder, Bagging estimator proved to be the best amongst the nine used ensemble learning models. Figure 18 summarizes the average F1 scores obtained by all ensemble models in the third study.

3) LANGUAGE MODELS RESULTS

Likewise, the same LLMs with the same hyperparameters used in the first study were used here. The BERT model achieved an F1-score of 0.44, 0.29, 0.42, and 0.42 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively with an average F1-score of 0.39. The RoBERTa model achieved an F1-score of 0.45, 0.32, 0.44, and 0.45 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively with an average F1-score of 0.42. OpenAI GPT model achieved an F1-score of 0.43, 0.30, 0.34, and 0.50 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively with an average F1-score of 0.39. The GPT2 model achieved an F1-score of 0.43, 0.41, 0.39, and 0.37 in the detection of ADHD, Anxiety, Bipolar, and Depression respectively with an average F1-score of 0.40. In this study, RoBERTa proved to be the best amongst the four LLMs used in the prediction of potential mental disorder. Figure 19 compares the average F1-scores obtained by the LLMs used in the third study.

At the end of this study, ML classifiers and ensemble learning models have outperformed Large Language Models (LLMs) in the task of the prediction of possible mental disorders from non-clinical text data or subreddits.

The champion model overall the ML classifiers is the logistic regression (LR) again, and the champion model overall the ensemble learning models used in this study is the Bagging estimator. Again, the champion models were chosen based on the average F1-score calculated from the F1-scores obtained for all disorders but the best model for every disorder is as follows: LR is the best in predicting depression, Bagging estimator is the best in predicting ADHD, and both models can equally predict possible anxiety and bipolar.

Table 1 summarizes the results obtained in every study of the three conducted studies which are: detection of mental disorder form clinical subreddits, detection of mental disorder

TABLE 2. Comparison between the proposed work and the state-of-the-art in terms of Dataset size, number of studied mental disorders, and number of models built and compared.

Authors	Year	Dataset Size	Number of addressed Mental disorders	No. of models	Best results
Kumar et al [8]	2019	100 users	1 (depression)	3 ML models	Acc= 85%
Tarik et al [9]	2019	Not mentioned	1 (depression)	2 DL models	Acc= 74%
Hussain et al [10]	2019	Not mentioned	2 (depression & anxiety)	3 ML models	F1-score= 0.84
Wong et al [9]	2019	Not mentioned	1 (depression)	2 DL models	-
Rezaii et al [11]	2019	40 users	1 (depression)	2 NLP techniques	Acc= 90%
Inkpen et al [12]	2019	Not mentioned	2 (depression and PTSD)	1 DL model	Acc= 88%
Thorstad et al [6]	2019	All REDDIT	4 (all mental disorders)	1 ML model	F1-score= 0.77
Trifan et al [13]	2020	Not mentioned	1 (depression)	3 ML models	F1-score= 0.72
Jiang et al [14]	2020	Not mentioned	4 (all mental disorders)	1 DL model	F1-score= 0.64
Alghamdi et al [15]	2020	Not mentioned	1 (depression)	6 ML models	Acc= 80%
Birnbaum et al [16]	2020	223 users	1 (depression)	2 ML models	Acc= 77%
Chatterjee et al [17]	2021	Not mentioned	1 (depression)	1 ML models	Acc= 76%
Ren et al [18]	2021	Not mentioned	1 (depression)	1 DL models	Acc= 91%
Shaoxiong et al [19]	2022	All REDDIT	1 (depression)	2 EL models	Acc= 75%
Nalini. L [20]	2022	Not mentioned	Not mentioned	3 ML models	Not mentioned
Tufail [21]	2023	Not mentioned	1 (depression)	1 DL model	Acc= 64%
Koushik et al [22]	2023	Not mentioned	1 (depression)	1 ML & 2 DL	Acc= 60%
Yicheng et al [24]	2023	Not mentioned	1 (depression)	1 Time series approach	-
Helmy et al [26]	2024	70,000 tweets	1 (depression)	5 ML models	Acc=92%
Dhariwal [27]	2024	Small healthcare dataset	1 (depression)	1 DL model	Acc=99.7%
This work	**2024**	**All REDDIT**	**4 (all mental disorders)**	**6 ML models 9 EL models 4 LLMs**	**F1-score= 0.80**

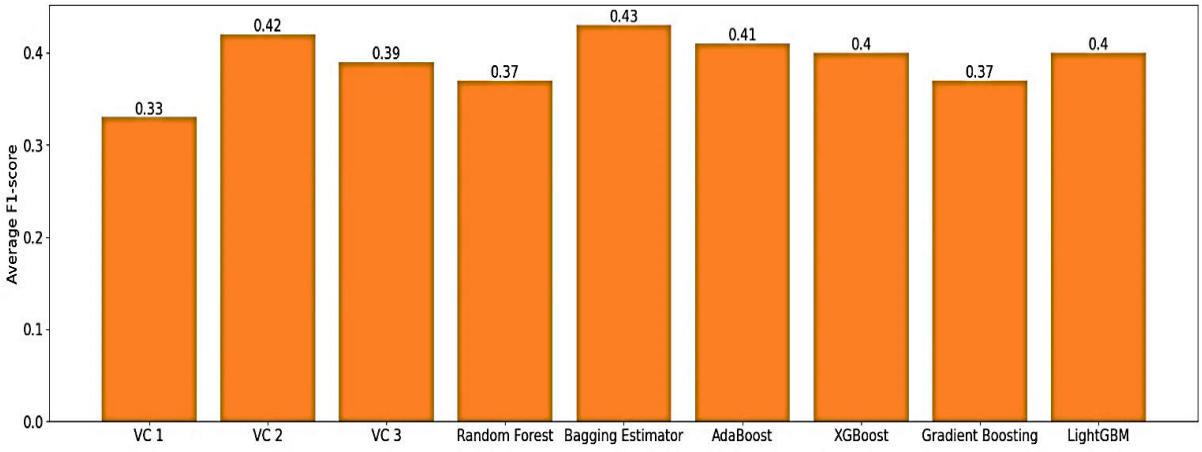

FIGURE 18. Study 3: F1-score for ensemble learning models.

TABLE 3. Comparison between the proposed work and the state-of-the-art in terms of results obtained in detection of every mental disorder from clinical subreddits (Study 1), with the best results highlighted using bold font.

Authors	Year	Methodology	ADHD	Anxiety	Bipolar	depression	Average
Thorstad et al [6]	2019	LR	0.83	0.75	0.75	0.74	0.77
This Work	2024	BERT	0.86	**0.79**	**0.79**	**0.76**	**0.80**
This Work	2024	RoBERTa	**0.87**	**0.79**	**0.79**	**0.76**	**0.80**
This Work	2024	Open AI GPT	0.86	0.78	**0.79**	**0.76**	**0.80**

from non-clinical subreddits, and the prediction of future possibility of mental dis0-order from non-clinical subreddits. Table 2 compares the work proposed in this study with all the previous work done in this area in terms of dataset size, number of studied mental disorders, and number of models built and compared.

Tables 3, 4, and 5 compare the proposed work and the most related state-of-the-art work in literature that addresses

TABLE 4. Comparison between the proposed work and the state-of-the-art in terms of results obtained in detection of every mental disorder from non-clinical subreddits (Study 2), with the best results highlighted using bold font.

Authors	Year	Methodology	ADHD	Anxiety	Bipolar	depression	Average
Thorstad et al [6]	2019	LR	0.42	0.30	0.34	0.38	0.38
This Work	2024	Bagging Est.	0.52	0.47	0.51	**0.57**	**0.52**
This Work	2024	XGBoost	**0.52**	0.47	**0.54**	0.53	**0.52**
This Work	2024	LightGBM	0.50	**0.49**	0.53	0.55	**0.52**

TABLE 5. Comparison between the proposed work and the state-of-the-art in terms of results obtained in prediction of every future possible mental disorder from non-clinical subreddits (Study 3), with the best results highlighted using bold font).

Authors	Year	Methodology	ADHD	Anxiety	Bipolar	depression	Average
Thorstad et al [6]	2019	LR	0.39	0.32	0.37	0.36	0.36
This Work	2024	Bagging Est.	**0.42**	**0.35**	**0.43**	0.50	**0.43**

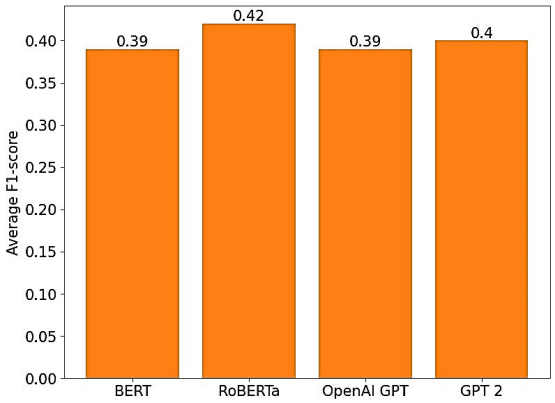

FIGURE 19. Study 3: F1-score for large language models.

the same problem [6]. The authors in literature used logistic regression to detect all types of mental disorder from social media posts using clinical subreddits and non-clinical subreddits. The authors also tried to predict the possibility of occurrence of all types of mental disorder from non-clinical subreddits. The approaches proposed in this work and the state-of-the-art approach used the same REDDIT to train and evaluate the suggested models.

VI. CONCLUSION

The proposed work aims at the early detection and even the prediction of potential future mental disorder from social media data. This successful approach can be used for effectively diagnosing mental disorders of social media users without asking them to cooperate in the diagnosis process. The successful diagnosis then can be further used to give advice or recommendations for early treatment and prevention of mental disorders.

In this work three different studies were conducted to cover all possible aspects in the field of analyzing social media posts to obtain statistical data about users' mental cases. The proposed work covers the four well-known mental disorders which are depression, anxiety, bipolar, and ADHD. The three different studies are: 1) mental disorder detection from clinical data, 2) mental disorder detection from non-clinical data which is a harder problem because no mental disorders are discussed or mentioned in this ordinary type of social media data, and 3) mental disorder prediction from non-clinical data which is further harder than the two previously mentioned studies.

REDDIT social media platform was used to train and evaluate the models used in this work, because it is the largest and most up-to-date publicly available social media data. With six classical machine learning classifiers, nine ensemble learning models, and four language models, this is the first study in literature that builds, trains, evaluates, and compares this large number of models to address mental disorder detection and prediction from social media data.

Large Language Models (LLMs) outperformed the classical machine learning classifiers, and the ensemble learning models in the task of detecting mental disorders from clinical data. The champion models overall the LLMs used in this task are BERT, RoBERTa, and OpenAI GPT. The champion models were chosen based on the average F1-score taken from the F1-scores obtained for every disorder, but some models were better than others per every disorder. BERT proved to be the best in detecting anxiety and bipolar, meanwhile RoBERTA proved to be the best in detecting ADHD, and OpenAI GPT proved to be the best in detecting depression (see table 1).

On the other hand, ML classifiers and ensemble learning models have outperformed Large Language Models (LLMs) in the tasks of detecting and predicting mental disorders from non-clinical data. The champion model overall the ML classifiers in both tasks is the logistic regression (LR), and the champion models overall the ensemble learning models used in the detection of already existing mental disorder are Bagging estimator, XGBoost, and LightGBM. LR and Bagging estimator are the best in detecting depression, XGBoost is the best in detecting anxiety and bipolar, and LightGBM is the best in detecting ADHD. The champion model that surpassed all the ensemble learning models used in the task of predicting future mental disorder is the Bagging estimator. LR is the best in predicting depression, Bagging estimator is the best in predicting ADHD, and both models can equally predict possible anxiety and bipolar (see table 1).

As a future work a comprehensive error analysis should be done to analyze the results obtained by every model especially the champion models to be able to further enhance the results of mental disorder detection and prediction either by changing the settings and hyperparameters of the models or by changing the data pre-processing part. Another important future work suggestion is the comprehensive fine-tuning of the hyper parameters which requires a lot of time and computational power, specially that feature selection step would be very important in that case to avid overfitting in such a large dataset. Cross validation should also be considered in the future work to validate the obtained results, validate the hyperparameters, finetune them based on validation results, and accordingly improve the testing results. Developing an end-to-end software tool which helps social media users with existing or potential mental disorders would be also a great added step to this work. Possible scenarios of the software tool could be a chatbot that helps mental disorder patients based on psychology knowledgebase. It could also be just a recommendation system that set the user aware of his/her case and give possible advice. It could be connected to psychology analysis exam to give more accurate diagnosis, and it might also be connected to contacts and data of psychiatrists recommended based on the case. To be able to deploy such a software tool, a quantization step for the LLMs would be of great added value to shrink the size of the models and accordingly make data handling easier.

ACKNOWLEDGMENT

The authors would like to thank Abdelrahman Hosny (an Assistant Lecturer with Assiut University), Mina Awad (an AI Innovation and Analytics Senior Manager), Hamis Hesham (a Senior AI and NLP Specialist), and Mahmoud Elsherif (AI and NLP Engineer) whom without them support granting them resources this work wouldn't be finished.

REFERENCES

[1] G. Scott, "Lifespan (half-life) of social media posts: Update for 2024," 2024, doi: 10.13140/RG.2.2.21043.60965.
[2] M. Qian and C. Kong, "Enabling human-centered machine translation using concept-based large language model prompting and translation memory," in *Proc. Int. Conf. Human-Comput. Interact.*, 2024, pp. 118–134.
[3] I. Frommholz, P. Mayr, G. Cabanac, and S. Verberne, "Bibliometric-enhanced information retrieval: 14th international BIR workshop (BIR 2024)," in *Proc. Eur. Conf. Inf. Retr.*, 2024, pp. 442–446, doi: 10.1007/978-3-031-56069-9_61.
[4] Q. Wang, "Text memorization: An effective strategy to improve Chinese EFL learners' argumentative writing proficiency," *Frontiers Psychol.*, vol. 14, Apr. 2023, Art. no. 1126194, doi: 10.3389/fpsyg.2023.1126194.
[5] N. Straková and J. Válek, "Chatbots as a learning tool: Artificial intelligence in education," *R E-Source*, pp. 245–265, Jan. 2024, doi: 10.53349/resource.2024.is1.a1259.
[6] R. Thorstad and P. Wolff, "Predicting future mental illness from social media: A big-data approach," *Behav. Res.*, vol. 51, pp. 1586–1600, Aug. 2019, doi: 10.3758/s13428-019-01235-z.
[7] [Online]. Available: https://www.reddit.com/dev/api
[8] A. Kumar, A. Sharma, and A. Arora, "Anxious depression prediction in real-time social data," in *Proc. Int. Conf. Adv. Eng., Sci., Manag. Technol.*, 2019, pp. 1–7.
[9] A. Wongkoblap, M. A. Vadillo, and V. Curcin, "Predicting social network users with depression from simulated temporal data," in *Proc. IEEE EUROCON-18th Int. Conf. Smart Technol.*, Jul. 2019, pp. 1–6.
[10] S. Tariq, N. Akhtar, H. Afzal, S. Khalid, M. R. Mufti, S. Hussain, A. Habib, and G. Ahmad, "A novel co-training-based approach for the classification of mental illnesses using social media posts," *IEEE Access*, vol. 7, pp. 166165–166172, 2019, doi: 10.1109/ACCESS.2019.2953087.
[11] N. Rezaii, E. Walker, and P. Wolff, "A machine learning approach to predicting psychosis using semantic density and latent content analysis," *Npj Schizophrenia*, vol. 5, no. 1, Jun. 2019.
[12] P. Kirinde Gamaarachchige and D. Inkpen, "Multi-task, multi-channel, multi-input learning for mental illness detection using social media text," in *Proc. 10th Int. Workshop Health Text Mining Inf. Anal. (LOUHI)*, Hong Kong, 2019, pp. 54–64.
[13] A. Trifan, R. Antunes, S. Matos, and J. L. Oliveira, "Understanding depression from psycholinguistic patterns in social media texts," in *Proc. Eur. Conf. Inf. Retr.*, vol. 12036, 2020, pp. 402–409.
[14] Z. Jiang, S. I. Levitan, J. Zomick, and J. Hirschberg, "Detection of mental health from Reddit via deep contextualized representations," in *Proc. 11th Int. Workshop Health Text Mining Inf. Anal.*, 2020, pp. 147–156.
[15] N. S. Alghamdi, H. A. Hosni Mahmoud, A. Abraham, S. A. Alanazi, and L. García-Hernández, "Predicting depression symptoms in an Arabic psychological forum," *IEEE Access*, vol. 8, pp. 57317–57334, 2020.
[16] M. L. Birnbaum, R. Norel, A. Van Meter, A. F. Ali, E. Arenare, E. Eyigoz, C. Agurto, N. Germano, J. M. Kane, and G. A. Cecchi, "Identifying signals associated with psychiatric illness utilizing language and images posted to Facebook," *Npj Schizophrenia*, vol. 6, no. 1, p. 38, Dec. 2020.
[17] R. Chatterjee, R. K. Gupta, and B. Gupta, "Depression detection from social media posts using multinomial naive theorem," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 1022, no. 1, 2021, Art. no. 012095, doi: 10.1088/1757-899X/1022/1/012095.
[18] L. Ren, H. Lin, B. Xu, S. Zhang, L. Yang, and S. Sun, "Depression detection on Reddit with an emotion-based attention network: Algorithm development and validation," *JMIR Med. Informat.*, vol. 9, no. 7, Jul. 2021, Art. no. e28754.
[19] L. Ansari, S. Ji, Q. Chen, and E. Cambria, "Ensemble hybrid learning methods for automated depression detection," *IEEE Trans. Computat. Social Syst.*, vol. 10, no. 1, pp. 211–219, Feb. 2023, doi: 10.1109/TCSS.2022.3154442.
[20] L. Nalini, "Methods in predictive techniques for mental health status on social networks using machine learning," *Int. J. Adv. Res. Sci., Commun. Technol.*, vol. 2, no. 1, Jul. 2022.
[21] H. Tufail, S. M. Cheema, M. Ali, I. M. Pires, and N. M. Garcia, "Depression detection with convolutional neural networks: A step towards improved mental health care," *Proc. Comput. Sci.*, vol. 224, pp. 544–549, Jan. 2023, doi: 10.1016/j.procs.2023.09.079.
[22] L. Koushik, M. A. Kumar, and R. L. Hariharan, "Interns@LT-EDI: Detecting signs of depression from social media text," in *Proc. 3rd Workshop Language Technol. Equality, Diversity Inclusion*, 2023, pp. 262–265, doi: 10.26615/978-954-452-084-7_040.
[23] K. M. Hasib, M. R. Islam, S. Sakib, Md. A. Akbar, I. Razzak, and M. S. Alam, "Depression detection from social networks data based on machine learning and deep learning techniques: An interrogative survey," *IEEE Trans. Computat. Social Syst*, vol. 10, pp. 1568–1586, 2023, doi: 10.1109/TCSS.2023.3263128.
[24] Y. Cai, H. Wang, H. Ye, Y. Jin, and W. Gao, "Depression detection on online social network with multivariate time series feature of user depressive symptoms," *Exp. Syst. Appl.*, vol. 217, May 2023, Art. no. 119538, doi: 10.1016/j.eswa.2023.119538.
[25] Z. Li, Z. An, W. Cheng, J. Zhou, F. Zheng, and B. Hu, "MHA: A multimodal hierarchical attention model for depression detection in social media," *Health Inf. Sci. Syst.*, vol. 11, no. 1, p. 6, Jan. 2023, doi: 10.1007/s13755-022-00197-5.
[26] A. Helmy, R. Nassar, and N. Ramdan, "Depression detection for Twitter users using sentiment analysis in English and Arabic tweets," *Artif. Intell. Med.*, vol. 147, Jan. 2024, Art. no. 102716, doi: 10.1016/j.artmed.2023.102716.
[27] N. Dhariwal, N. Sengupta, M. Madiajagan, K. K. Patro, P. L. Kumari, N. A. Samee, R. Tadeusiewicz, P. Plawiak, and A. J. Prakash, "A pilot study on AI-driven approaches for classification of mental health disorders," *Frontiers Hum. Neurosci.*, vol. 18, Apr. 2024, Art. no. 1376338, doi: 10.3389/fnhum.2024.1376338.
[28] Z. Wu, Z. Wang, J. Chen, H. You, M. Yan, and L. Wang, "Stratified random sampling for neural network test input selection," *Inf. Softw. Technol.*, vol. 165, Jan. 2024, Art. no. 107331, doi: 10.1016/j.infsof.2023.107331.

[29] J. A. Adeyiga, P. G. Toriola, T. E. Abioye, and A. E. Oluwatosin, "Fake news detection using a logistic regression model and natural language processing techniques," 2023, doi: 10.21203/rs.3.rs-3156168/v1.

[30] B. M. Iqbal, K. M. Lhaksmana, and E. B. Setiawan, "2024 presidential election sentiment analysis in news media using support vector machine," *J. Comput. Syst. Informat. (JoSYC)*, vol. 4, no. 2, pp. 397–404, Feb. 2023, doi: 10.47065/josyc.v4i2.3051.

[31] Y. Zhang, H. Li, Z. Li, N. Cheng, M. Li, J. Xiao, and J. Wang, "Leveraging biases in large language models: 'Bias-kNN' for effective few-shot learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024.

[32] I. H. Sodhar and A. H. Buller, "Natural language processing: Applications techniques and challenges," 2020, doi: 10.22271/ed.book.784.

[33] Y. Tian, Y. Zhang, and H. Zhang, "Recent advances in stochastic gradient descent in deep learning," *Mathematics*, vol. 11, no. 3, p. 682, Jan. 2023, doi: 10.3390/math11030682.

[34] A. H. Odeh, M. Odeh, and N. Odeh, "Using multinomial naive Bayes machine learning method to classify, detect, and recognize programming language source code," in *Proc. Int. Arab Conf. Inf. Technol. (ACIT)*, Abu Dhabi, United Arab Emirates, Nov. 2022, pp. 1–5, doi: 10.1109/ACIT57182.2022.9994117.

[35] L. Geni, E. Yulianti, and D. I. Sensuse, "Sentiment analysis of tweets before the 2024 elections in Indonesia using IndoBERT language models," *J. Ilmiah Teknik Elektro Komputer Dan Informatika*, vol. 9, pp. 746–757, Sep. 2023.

[36] I. Setiawan, A. M. Widodo, M. Rahaman, M. A. Hadi, N. Anwar, M. B. Ulum, E. Y. Mulyani, and N. Erzed, "Utilizing random forest algorithm for sentiment prediction based on Twitter data," in *Proc. 1st Mandalika Int. Multi-Conf. Sci. Eng.*, 2022, pp. 446–456, doi: 10.2991/978-94-6463-084-8_37.

[37] D. Tiwari, B. Nagpal, B. S. Bhati, A. Mishra, and M. Kumar, "A systematic review of social network sentiment analysis with comparative study of ensemble-based techniques," *Artif. Intell. Rev.*, vol. 56, no. 11, pp. 13407–13461, Nov. 2023.

[38] Q. Sui and S. K. Ghosh, "Active learning for stacking and AdaBoost-related models," *Stats*, vol. 7, no. 1, pp. 110–137, Jan. 2024, doi: 10.3390/stats7010008.

[39] G. Hu, M. Ahmed, and M. R. L'Abbé, "Natural language processing and machine learning approaches for food categorization and nutrition quality prediction compared with traditional methods," *Amer. J. Clin. Nutrition*, vol. 117, no. 3, pp. 553–563, Mar. 2023, doi: 10.1016/j.ajcnut.2022.11.022.

[40] A. Villar and C. R. V. de Andrade, "Supervised machine learning algorithms for predicting student dropout and academic success: A comparative study," *Discover Artif. Intell.*, vol. 4, no. 1, pp. 1–24, Jan. 2024, doi: 10.1007/s44163-023-00079-z.

[41] B. Abu-Salih, S. Alotaibi, R. Abukhurma, M. Almiani, and M. Aljaafari, "DAO-LGBM: Dual annealing optimization with light gradient boosting machine for advocates prediction in online customer engagement," *Cluster Comput.*, pp. 1–27, 2024, doi: 10.1007/s10586-023-04220-6.

[42] M. Islam and L. Zhang, "A review on BERT: Language understanding for different types of NLP task," 2024, doi: 10.20944/preprints202401.1857.v1.

[43] A. Sobhy, M. Helmy, M. Khalil, S. Elmasry, Y. Boules, and N. Negied, "An AI based automatic translator for ancient hieroglyphic language—From scanned images to English text," *IEEE Access*, vol. 11, pp. 38796–38804, 2023, doi: 10.1109/ACCESS.2023.3267981.

[44] N. A. Semary, W. Ahmed, K. Amin, P. Plawiak, and M. Hammad, "Improving sentiment classification using a RoBERTa-based hybrid model," *Frontiers Human Neurosci.*, vol. 17, Dec. 2023, Art. no. 1292010, doi: 10.3389/fnhum.2023.1292010.

[45] N. K. Negied, S. H. Anwar, K. M. Abouaish, E. M. Matta, A. A. Ahmed, and A. K. Farouq, "Academic assistance chatbot—A comprehensive NLP and deep learning-based approaches," *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 33, no. 2, p. 1042, Feb. 2024, doi: 10.11591/ijeecs.v33.i2.pp1042-1056.

[46] J. Choi and B. Lee, "Accelerated materials language processing enabled by GPT," 2023, *arXiv:2308.09354*.

[47] K. Y. Thakkar and N. Jagdishbhai, "Exploring the capabilities and limitations of GPT and chat GPT in natural language processing," *J. Manage. Res. Anal.*, vol. 10, no. 1, pp. 18–20, Apr. 2023, doi: 10.18231/j.jmra.2023.004.

[48] X. Zheng, C. Zhang, and P. C. Woodland, "Adapting GPT, GPT-2 and BERT language models for speech recognition," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Cartagena, Colombia, Dec. 2021, pp. 162–168, doi: 10.1109/ASRU51503.2021.9688232.

[49] O. O. Awe, G. O. Opateye, C. A. G. Johnson, O. T. Tayo, and R. Dias, "Weighted hard and soft voting ensemble machine learning classifiers: Application to anemia diagnosis," in *Sustainable Statistical and Data Science Methods and Practices*, 2024, doi: 10.1007/978-3-031-41352-0_18.

MOHAMMED ABDULLAH received the bachelor's degree in computers and information (majoring in information systems) from Assiut University, Egypt, in June 2019, and the master's degree in electrical and computer engineering from the University of Ottawa, Canada, in March 2023.

He is an Natural Language Processing (NLP) Engineer. He has a strong background in machine learning, deep learning, and NLP. His research interests and areas of expertise include NLP in Arabic and English, where he employs cutting-edge machine learning and deep learning methods.

NERMIN NEGIED (Member, IEEE) received the M.Sc. and Ph.D. degrees from the Computer Department, Faculty of Engineering, Cairo University, in February 2012 and July 2016, respectively. She is the Head of data science and artificial intelligence track with the Digital Egypt Builders Initiative (DEBI), Egyptian Ministry of Communication and Information Technology (MCIT). She is an Assistant Professor with Cairo University; New Giza University; and the Faculty of Engineering and Computer Science, Nile University. She is a former Assistant Professor with Zewail City of Science and Technology, Arab Academy for Science and Technology and Maritime Transport (AASTMT), and October University for Modern Science and Arts (MSA). She was the Educational Quality Manager with the Faculty of Computer Science, MSA. She was a Lecturer with the Computer Engineering Department, Faculty of Engineering, 6th of October University, from September 2012 to September 2015, where she was a Teaching Assistant, from September 2006 to September 2012. She has published several international journals and conference papers and shared in reviewing several scientific papers. Her research interests include image processing and computer vision, machine learning, artificial intelligence, expert systems, natural language processing, and LLMs.