IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 June 2024, accepted 14 June 2024, date of publication 24 June 2024, date of current version 1 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3417821

RESEARCH ARTICLE

Toward Early Detection of Depression: Detecting Depression Symptoms in Arabic Tweets Using Pretrained Transformers

SUZAN ELMAJALI[1] AND IRFAN AHMAD[1,2]
[1]Department of Information and Computer Science, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia
[2]SDAIA–KFUPM Joint Research Center for Artificial Intelligence, KFUPM, Dhahran 31261, Saudi Arabia

Corresponding author: Irfan Ahmad (irfan.ahmad@kfupm.edu.sa)

This work was supported by the Saudi Data and Artificial Intelligence Authority (SDAIA) and KFUPM through the SDAIA-KFUPM Joint Research Center for Artificial Intelligence under Grant JRC–AI–RFP–10.

ABSTRACT The COVID-19 pandemic and its associated setbacks have significantly impacted human mental health. Depression of various intensities has resulted due to a wide variety of losses that people have experienced. However, unlike physical illness, mental illness is still underestimated by patients themselves and by society due to various factors, such as the societal stigma of visiting a psychotherapist and being diagnosed with a mental health disorder. On the other hand, general practitioners can recognize signs of depression using the Patient Health Questionnaire (PHQ-9), which is used as a screening test for depression. The PHQ-9 questionnaire comprises nine questions that correspond to the nine symptoms of depression outlined in the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5). In this paper, we aim to detect the nine depression symptoms stated by DSM-5 from Arabic tweets, as recognizing the type of depression symptom is crucial in diagnosing depression. We used AraBERT and MARBERT pretrained transformers to classify tweets with depression symptoms. We also performed data augmentation using ChatGPT to balance the training set. The model was applied to a dataset consisting of 1,290 samples labeled with nine different symptoms, in addition to a 'normal' class which was also generated using ChatGPT. This work used four performance metrics to evaluate the models' performance, which are accuracy, precision, recall, and F1 scores. The AraBERT and MARBERT transformers have yielded promising results, achieving accuracy, precision, recall, and F1 scores of 99.3%, 99.1%, 98.8%, and 98.9%, respectively, using the AraBERT transformer. While using the MARBERT transformer achieved accuracy, recall, precision, and F1 scores of 98.3%, 97.9%, 98.2%, and 98%, respectively.

INDEX TERMS Arabic tweets, depression detection, model finetuning, text classification, transformer models.

I. INTRODUCTION

Recently, concerns about human mental health have been raised, especially after the COVID-19 outbreak in 2019. COVID-19 has had a negative impact on many people's mental health, leading to various psychological disorders such as depression, worry, and anxiety due to the unusual situation they were experiencing [1]. Even after the COVID-19, people may still face several circumstances which may lead to psychological disorders, like social reasons, work pressure, and economic challenges, which means the need for psychotherapists is increasing. On the other hand, many individuals still consider reaching psychotherapy clinics as a stigma [2], which is considered an obstacle for many people to reach psychotherapists when they face any mental disorders. Nevertheless, Artificial Intelligence (AI) solutions started to appeal to give a hand in the awareness and treatment of depression patients [3], [4]. Moreover, social

The associate editor coordinating the review of this manuscript and approving it for publication was Muammar Muhammad Kabir.

media content is being used as an effective component to detect depression as in [5], and this is due to many people feel comfortable expressing their feelings and opinions on social media behind the screens. However, due to the availability of English datasets, many of the proposed applications for depression detection were focused on English speakers, like the psychotherapist chatbot ''Weobot'' which has shown interesting results in overcoming some mental disorders like anxiety and depression [6]. In addition to other proposed applications to detect depression from English text like [7]. To diagnose a person with depression, physicians usually start diagnosing using a screening tool called Patient Health Questionnaire-9 (PHQ-9) [8]. However, it is important to note that this tool is not the only tool used in depression diagnostic. Rather, it is considered as an indicator tool which shows if the person may need medical intervention and treatments. Fig. 1 illustrates the PHQ-9 questions, which correspond to the nine symptoms of depression.

The PHQ-9 questionnaire is composed of nine questions, each representing the presence of a specific symptom of depression symptoms which are outlined in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [9]. When multiple symptoms appear in a patient's life within the past 2 weeks, the test score increases. If the score reaches 10 or higher, medical intervention may be required. Therefore, it is crucial to distinguish between these symptoms to achieve a more precise diagnostic process. So for example, if a person has been experiencing only sleep disorders for the past few weeks without exhibiting any of the other symptoms among the nine depression symptoms, we can not consider it as an indication of depression. So Each symptom of the nine depression symptoms must be identified and differentiated from other symptoms to obtain a more accurate diagnostic process.

Below, we will provide explanations for the nine symptoms, accompanied by relevant words and phrases that serve as indicators for each symptom.

- Losing interest or pleasure in activities: where an individual lacks excitement or pleasure in engaging in tasks they once found enjoyable. This symptom can be identified through the use of words or phrases like boring, uninteresting, and losing passion.
- Low mood: this refers to a prolonged state characterized by feelings of sadness and unhappiness. The symptom can be expressed through words such as ''sad'', ''unhappy'', ''cry''.
- Sleep disorder: experiencing a disruption in sleep patterns and a persistent inability to sleep, without any discernible reasons. On the other hand, the case can be the opposite by having excessive sleeping. This symptom can be characterized by phrases such as ''difficulty falling asleep,'' ''reluctance to wake up,'' or by using the term ''insomnia.''
- Weight disorder: this refers to fluctuations in body weight caused by eating disorders that arise from symptoms related to bad mood, stress or any negative feelings. This symptom can be characterized by phrases such as ''weight loss'', ''weight gain'' or ''loss of appetite.''
- Loss of energy: when a person has feelings of fatigue and a lack of physical or mental vitality. So they do not want to get engaged in any activity or perform any daily tasks. Words that may lead to this symptom are ''fatigue'' and ''exhaustion''.
- Feelings of worthlessness: this emotional state arises following a specific event that results in disappointment or feeling let down. Words such as ''disappointed,'' ''let down,'' and ''broken'' often indicate the emergence of this symptom.
- Diminished ability to think or concentrate: when an individual becomes unable to maintain focus and concentration on a particular task or activity due to being consistently distracted by their thoughts and emotions. Words and phrases attached to this symptom are: dispersed, can not focus, can not concentrate.
- Psychomotor agitation or retardation: refers to a state in which an individual experiences an uncontrollable urge to continuously move or engage in aimless physical actions, without a clear purpose or intention. This can manifest as restlessness, an inability to remain still, or slowed movements. This symptom can be characterized using words like ''laziness'' and ''lethargy''.
- Suicidality: the feeling or thoughts someone has of intentionally ending their own life. Suicidality symptom can manifest in a person's speech when they express desires to die or a profound sense of not wanting to continue living. This symptom can be described using phrases that indicate thoughts of death and desire to die.

Our objective, however, is to differentiate between symptoms in order to track them more effectively. In this paper, we aim to employ multi-class classification techniques to detect the nine depression symptoms from tweets. We aim to classify depression-related tweets into one of the nine specific depression symptoms and classify non-depression tweets as normal, hence automating the process of detecting and diagnosing depression from Arabic texts. This will facilitate the process of labelling huge Arabic datasets with depression and non-depression labels so that they can be used for other applications. Moreover, this will open the opportunity of achieving more accurate results in detecting depression in its early stages. The main contributions of our work are:

- We present automatic detection of nine depression symptoms from Arabic text. According to the best of our knowledge, this is the first work on depression symptom detection in Arabic.
- Customize Modern Standard Arabic Mood Changing and Depression dataset [10] by adding ''Normal'' class to it using ChatGPT, thus leading to a dataset consists of 10 classes.
- Applying data augmentation to the dataset considering the different Arabic dialects used in the dataset, using ChatGPT.

PHQ-9 depression questionnaire

Name:	Date:			
Over the last 2 weeks, how often have you been bothered by any of the following problems?	**Not at all**	**Several days**	**More than half the days**	**Nearly every day**
Little interest or pleasure in doing things	0	1	2	3
Feeling down, depressed, or hopeless	0	1	2	3
Trouble falling or staying asleep or sleeping too much	0	1	2	3
Feeling tired or having little energy	0	1	2	3
Poor appetite or overeating	0	1	2	3
Feeling bad about yourself, that you are a failure, or that you have let yourself or your family down	0	1	2	3
Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
Moving or speaking so slowly that other people could have noticed; or the opposite, being so fidgety or restless that you have been moving around a lot more than usual	0	1	2	3
Thoughts that you would be better off dead or of hurting yourself in some way	0	1	2	3
Total __ =	__	+ __	+ __	+ __
PHQ-9 score ≥10: Likely major depression				
Depression score ranges:				
5 to 9: mild				
10 to 14: moderate				
15 to 19: moderately severe				
≥20: severe				
If you checked off any problems, how difficult have these problems made it for you to do your work, take care of things at home, or get along with other people?	Not difficult at all __	Somewhat difficult __	Very difficult __	Extremely difficult __

PHQ: Patient Health Questionnaire.

Developed by Drs. Robert L Spitzer, Janet BW Williams, Kurt Kroenke, and colleagues, with an educational grant from Pfizer, Inc. No permission required to reproduce, translate, display or distribute.

Graphic 59307 Version 13.0

FIGURE 1. Patient Health Questionnaire-9 [8].

The rest of the paper is organized as follows: in Section II, we will go through the previous works which have been done in detecting depression from Arabic texts. In Section III, we will explain the methodology used in this work. In Section IV, we will evaluate and discuss the results achieved from the proposed models. Finally, we will end this paper with a discussion and conclusions about this work in Section V.

II. RELATED WORK

During the past several years, much research has been conducted about the ability of deep learning and natural language processing tools to detect depression and suicidal intentions from social- media posts. Some proposed studies primarily conducted binary classification by classifying posts as depression or normal, as in [11], The researchers conducted a binary classification task to detect depression from Arabic posts. They utilized two different Bidirectional Encoder Representations from Transformers(BERT) models, namely ARABERT and MARBERT. To train and evaluate these models, they created a balanced dataset consisting of 7000 normal and depressed posts. Instead of detecting depression based on symptom tracking, the researchers collected data for their dataset by scraping mental health websites and specifically identifying depressed posts. They also utilized Twitter data and consulted with psychiatrists to identify specific words associated with depression in order to collect depressed tweets. Therefore, the classification of depression was based on the content of the collected tweets rather than tracking symptoms. The ARABERT model achieved an accuracy of 96.93%, while the MARBERT model achieved an accuracy of 96.07%. while the results of the study showed promising performance and indicated that the BERT models outperformed other machine learning models, it is crucial to mention that the comparison with machine learning models in this study was conducted with different datasets. In [12] The authors conducted a comparison between convolution Neural Networks (CNN), Supported Vector Machine (SVM), and Long Short Term Memory (LSTM) models for the binary classification of depression and non-depression tweets. They utilized a dataset comprising 18,000 tweets that were manually labelled. After preprocessing the dataset, the CNN model demonstrated the best performance, achieving an accuracy of 88%. However, it is important to note that the study solely relied on word embeddings as inputs without incorporating any

additional features. Consequently, the SVM model exhibited the poorest performance among the models examined. A paper was proposed by Almouzini et al. about detecting depressed users from Arabic tweets [13] also conducted a binary classification (depression and healthy). The data was collected from 97 Twitter users, 35 users of them with depression. The authors extracted the history of all users (with depression and without depression) and finally labelled the data manually based on the PHQ-9 scale. The authors of this paper used four machine learning models (Random Forest(RF), Naïve Bayes, AdaBoostM1, and LibLinear) to conduct a binary classification. However, before applying the classifier, they extracted the most crucial features from the text to distinguish between classes. They constructed a set of unigrams and a set of negations about emotional expressions, depression synonyms and emojis. The results showed that Liblinear achieved the best performance with an average precision and recall of 87.6% and 87.5%. On the other hand, the poorest performance was by the AdaBoost model, as it achieved 55.2% accuracy. Alghamdi et al. used different types of texts and proposed a model to detect depression based on a corpus text rather than a single-line tweet [14]. As the corpus might hold several symptoms. In the proposed paper, the authors collected the corpus from Nafsany.com, a platform specialized for mental health disorders. Then, the data was labelled by an expert psychotherapist into two main labels: depression and non-depression. After cleaning and processing the data, they used three different approaches the first was the lexicon-based approach, where they created a lexicon of the depression-related terms. The lexicon was created by applying the n-gram models (unigram, bigram and trigram) on the depression and non-depression data and then calculating the frequency of each n-gram created token and taking the top 1000 grams, which are the most frequent lexicon. The second approach was a rule-based algorithm by taking the lexicon, which was created in the lexicon-based approach and extracting the best combinations of the grams. The output dataset from the previous approach was input into the rule-based algorithm and the lexicon file. The rule-based file classified the data into depression and non-depression based on a defined threshold. The third approach was about applying four different Machine Learning (ML) models (ADA boost, K-Nearest Neighbor (KNN), Decision Tree(DT), Random Forest(RF), SVM and Stochastic Gradient Descent(SGD). The lexicon-based approach has reached 80.45% accuracy. Regarding the machine learning approach, the best performance was by using SGD learning algorithm, which resulted in 73% accuracy. Reference [15] also conducted a binary classification on a dataset consisting of 6307 Arabic tweets with Saudi dialect labelled with normal and depressed. They used Term Frequency-Inverse Document Frequency (TF-IDF) and N-gram for feature extraction, then applied 4 different ML models (Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression (LR) and Naïve Bayes (NB))on the extracted features. LR achieved the best result with an accuracy of 82%. From the conducted experiments the TF-IDF positively affected the results when it was added to N-gram, however, still the dataset consists of one dialect which can affect the results when applied to Arabic tweets but with different dialects.

Suicidal intention is one of the depression symptoms, some proposed studies focused on detecting this symptom from Arabic tweets. Reference [16] proposed a model to automatically detect suicidal intention tweets, they created a novel Arabic suicidal dataset and applied several machine learning (ML) models on word frequency and embedding features. The results showed that the best performance was achieved using SVM with an accuracy of 86% and F1 79%. On the other hand, after conducting experiments using the pretrained model AraBERT, it outperformed all other ML models and achieved 91% accuracy and 88% F1. Reference [17] also proposed a model to classify tweets into two main categories (normal and suicide) as suicidal text is one of the 9 depression symptoms. The dataset utilized was obtained by scraping Twitter, resulting in a collection of 14,576 tweets. To preprocess the data, stemming and lemmatization techniques were applied. Then the authors conducted several experiments using different versions of Arabic BERT models like AraBERT and AraElectra, and different versions of universal-sentence-encoder-multilingual (USE) as they support the Arabic language. the highest performance was observed when utilizing USE-MQA on the unnormalized text, resulting in an accuracy of 83.3%. On the other hand, for BERT with the ''Before Normalization'' dataset, the best-achieved results reached an accuracy of 95.26% when employing AraBERT.

On the other hand, some proposed work in depression detection did three-class classification as in [18]. The authors did 3 classes classification. They classified the tweets into (depression, non-depression and neutral) where depression tweets hold at least one depression symptom while non-depression tweets don't hold any symptom and neutral tweets do not fall in any of the previous categories like songs or quotes. The proposed model started with data preprocessing like stemming and eliminating replicated characters, then extracting appropriate features. N-grams and term frequency-inverse document frequency (TF-IDF) were utilized. These features were subsequently inputted into six different supervised machine learning models: Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), Logistic Regression (LR), Adaboost, and Naive Bayes (NB). The results indicated that the RF model, utilizing TF-IDF features, achieved the best performance among the models evaluated, achieving 82.39% accuracy. Moreover, Hassib et al. did 3 classes classification using variant Arabic-based transformers to detect depression, suicidal ideation and non-depression from Arabic tweets [19]. They created a dataset consisting of 20,213 tweets (5,472 depression, 2,167 suicidal ideation, and 12,574 as non-depression) tweets were labelled manually. Furthermore, the researchers augmented their dataset by incorporating publicly available

Arabic mood-changing and depression datasets. After pre-processing it, 30 different Arabic-based transformers were applied to this dataset, Like ARAELECTRA, ARABERT and MARABERT. The highest results were achieved using MARBERT using the same dataset, as it achieved an 88.75% F1 score. On the other hand, the lack of available datasets that consider all nine symptoms may be one of the reasons why experiments in detecting depression often fall short of reflecting reality. However, there have been notable efforts to create datasets in the field of sentiment analysis, such as the dataset proposed by [20] for a dataset consisting of 20,000 annotated samples of Saudi dialect tweets labelled with positive and negative. Moreover, some proposed studies have focused on developing methods for augmenting the available dataset, as in [21] the authors proposed three different approaches for data augmentation for mental health classification on social media. The first approach was easy data augmentation by making synonym replacement, random insertion, random swap and deletion texts. The second approach was about using the fine-tuned BERT model, which they called AugBERT, to generate augmented texts. The third approach was about back translation; by changing the language of the textual data to another language and then translating it back to the original language, the generated text from this translation was used for data augmentation. The three approaches were conducted on English tweets from a dataset of 1895 tweets labelled ''depression'' and ''suicide''. The AugBERT method, which is the second method, significantly improved the model's performance by increasing the accuracy results from 0.66 to 0.71.

Moreover, some studies were proposed to detect depression from English text and did binary classification as well by classifying the text as depressed and normal. for example, [22] proposed the FCL model to detect depression from social media text. The proposed model used a combination of Fast text embedding technique along with Convolutional Neural Network(CNN) and Long Short Term Memory(LSTM) models. The model was implemented on two datasets. A large dataset comprising 13,000 records was gathered from Reddit and labelled with depression and non-depression. The second dataset is a smaller one, with more than 6100 tweets labelled with depression and N-DEPR. The model achieved 0.87 accuracy with the large dataset and 0.88 with the second dataset (the small dataset). The proposed model outperforms existing methods, such as a model that uses the word2vec embedding technique instead of fast text embedding. In a context closely related to depression detection, which is suicidal detection from tweets, the authors in [23] considered the history of tweets that the same person posted and applied an transformer-based model (Suicidality assessment Time aware Temporal Network) STATNet, to detect suicidal intent from English tweets using the history of the posted tweets. They used a dataset of over 34,000 tweets labelled as Suicidal Intent (SI) present and Suicidal Intent (SI) absent. The proposed model consisted of Individual Tweet Modeling (ITM), which uses a BERT transformer and Historic Tweet Modeling (HTM), using a Pluchtek Emonet transformer and LSTM layer. the proposed model achieved 0.85 accuracy which outperformed the contextual CNN model because of its ability to take longer histories and represent them more efficiently.

Based on the studied literature review, The proposed models tend to manually label the available texts as depression and non-depression by experts based on reading the text and detecting at least five symptoms from the nine depression symptoms, or labelling each depression symptom as depression. However, in order to bring the task of detecting depression from tweets closer to reality, we propose a different approach. Instead of focusing on detecting depression from the input tweet, we suggest detecting individual symptoms of depression, by classifying text to one of the 9 symptoms defined by PHQ-9. To the best of our knowledge, previous works have not explored this approach of symptom detection from Arabic Tweets. By detecting symptoms, we can lay a strong foundation for systems that can track the occurrence of symptoms over time and provide a depression diagnosis once a specific number of symptoms have been detected.

III. METHODOLOGY

In this section, we will go through the methodology which was followed in this work. We used AraBERT [24] and MARBERT [25] pre-trained models to classify Arabic tweets into ''normal'' or one of the nine depression symptoms, which are stated in PHQ-9 questioner and was explained in the introduction section: diminished ability to think or concentrate, feelings of worthlessness, losing interest or pleasure in activities, loss of energy, low mood, psychomotor agitation or retardation, sleep disorder, suicidality, and weight disorder. So the final number of classes in this classification task is 10 classes.

We ran the experiments using Google Colab GPU. Moreover, in this work, we used PyTorch, SKlearn, seaborn and nltk libraries in addition to the transformer library to apply the pre-trained models AraBERT and MARBERT.

A. AraBERT AND MARBERT MODELS

AraBERT and MARBERT are Arabic transformers based on the Bidirectional Encoder Representation from Transformer BERT [26]. These models have undergone extensive training on a large volume of Arabic data, including various dialects. However, given the relatively limited size of the dataset we will be working with, we will utilize these pre-trained models for our classification task. Additionally, according to existing literature, these models have shown promising results in the classification tasks of Arabic Tweets.

BERT, originally proposed by Google in 2019, serves as the foundation for AraBERT and MARBERT models. BERT model consists of 12 layers, 768 hidden units, and 12 heads was trained on unlabelled text, and was trained using two training approaches, next sentence prediction and masked language modelling. One of the key features of the BERT

TABLE 1. Samples from modern standard arabic mood changing and depression dataset.

Tweet	Label
اسبوع فيني خمول	Psychomotor agitation
عقلي مشتت	Diminished ability to think
خاب ظني	Feelings of worthlessness
ملل مش طبيعي	Losing interest
متعبة جداً	Loss of energy
احس بضيق الدنيا كلها قلبي	Low mood
ابدا مو قادره انام	Sleep disorder
ابي اموت	Suicidality
مالي نفس اكل	Weight disorder

model is its ability to capture the context of the text which made it state of the art in some tasks like text classification.

AraBERT is a pre-trained model that consists of 12 attention layers, 12 attention heads, 768 hidden dimensions, and a maximum sequence length of 512. It was trained using two approaches: next-sentence prediction and masked language modelling, following the same methodology as the BERT model. Specifically designed for the Arabic language, AraBERT was trained on 2.5 billion Arabic tokens sourced from various domains, with a predominant usage of Modern Standard Arabic (MSA). This extensive training enables the model to comprehend different Arabic dialects and styles, thereby enhancing its efficacy in text classification tasks.

On the other hand, MARABERT is similar to AraBERT, which is a pre-trained model that was constructed and trained using a methodology similar to BERT. This involved employing next-sentence prediction and masked language modelling techniques. However, what sets MARABERT apart from AraBERT is its training dataset, which is specifically focused on the dialectal Arabic corpus. Additionally, MARABERT has a larger number of tokens and vocabulary size, with a vocabulary of 2.5 billion and 15.6 tokens.

B. DATASET

In this work, we used Modern Standard Arabic Mood Changing and Depression Dataset [10]. The data consists of 1230 samples labelled with nine depression symptoms. The samples are Arabic tweets which were automatically collected using TWINT library. The keywords used for collecting these tweets were related to the depression symptoms mentioned in the PHQ-9 questionnaire. At the first step of collecting the data, the retrieved data was around 48 thousand records. Most of them were eliminated and removed as duplicate tweets or irrelevant tweets.

The retrieved data was processed and cleaned by removing urls, duplicate letters, usernames, punctuation, stop words, emojis, and English words. In addition, the text was normalized by removing Altashkeel and elongation. The final dataset consists of 1230 Arabic tweets. A sample of the tweets is shown in Table 1.

To analyse the dataset and get a quick summary of the most common words in the dataset, a word cloud as in Fig. 2 was generated.

C. PREPROCESSING AND ADDING ADDITIONAL CLASS TO THE DATASET

Although the data was preprocessed and cleaned, it still needs some preprocessing before inputting it into the models. The first step in preprocessing was removing empty rows in the dataset as there were around 7 empty rows, so the size of the dataset became 1223. After removing the empty rows, we converted the labels to numeric values as it will facilitate working on the data. The conversion was by representing the first symptom which is ''diminished ability to think or concentrate'', as 0, the second symptom ''feelings of worthlessness'' as 1, the third symptom ''losing interest or pleasure in activities'' as 2, the fourth symptom ''loss of energy'' as 3, the fifth symptom ''low mood'' as 4, the sixth symptom ''psychomotor agitation or retardation'' as 5, the seventh symptom ''sleep disorder'' as 6, the eighth symptom ''suicidality'' as 7 and the ninth symptom ''weight disorder'' as 8.

Nevertheless, since the dataset has nine classes representing the nine different depression symptoms, we decided to add class 10 to the dataset, which represents ''normal'' tweets and represented numerically as 9. Tweets which belong to class ''normal'', are tweets that do not hold any depression symptoms in their text, as adding this class will train the model on classifying the normal cases where the text does not hold any depression symptoms. Adding the 10th class was done using ChatGPT by inputting one tweet from each class in the dataset and requesting a text using the same dialects but holding the opposite feeling of the input tweet. The following list shows a sample of sentences we got from ChatGPT when we input a tweet which says

سلاما لمن يجلس وحيدا مشتت لافكار مبعثر المشاعر تماما مثلي

which was selected from ''diminished ability to think or concentrate'' class:

1. انا سعيد جداً بأنني لست وحيداً اليوم
2. لدي الآن الوقت الكافي لترتيب أفكاري ومشاعري
3. سأذهب لأرى أصدقائي المقربين اليوم و أشعر بالسعاده
4. أحب الجلوس وحدي في الهواء الطلق و التأمل في الطبيعة الجميلة
5. سوف أشاهد فيلمًا جيدًا وأستمتع بوقتي اليوم
6. لدي الكثير لأشكر الله عليه في حياتي، وسأركز على الإيجابيات اليوم

For each input tweet, we requested 20-30 different sentences holding the opposite feeling of the tweet's text. Ultimately, we got a ''Normal'' class with 250 samples resulting in a total of 1471 samples within the dataset. Fig. 3 indicates the dataset distribution over ten classes. The plot indicates clearly that the classes are not balanced, with a higher number of records in the classes ''diminished ability to think or concentrate,'' ''weight disorder,'' and ''Normal'' compared to the other classes. So in order to make

FIGURE 2. Word Cloud for 1230 Arabic tweets in Modern Standard Arabic Mood Changing and Depression Dataset.

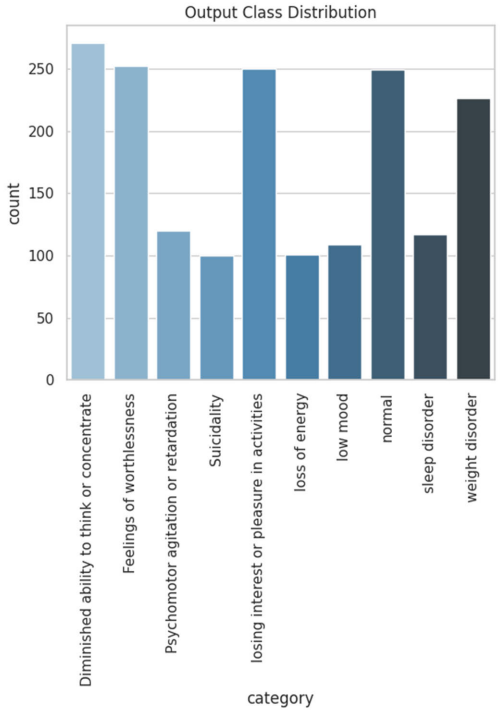

FIGURE 3. Dataset distribution over 10 classes.

them balance, a data augmentation was implemented on the training data after splitting the data. The details of the splitting and augmentation process will be discussed in the subsequent sections.

D. SPLITTING THE DATASET

The data was split into training, validation, and testing sets. First, we split the data into training and testing sets, with a percentage of 80% for the training data and 20% for the testing data, this resulted in 295 samples for testing and 1176 samples for training. Then, we did data augmentation on the training data, which will be described in detail in the following subsection. Subsequently, the augmented training data was further split into training and validation sets, with percentages of 80% and 20%, respectively.

When splitting the data, we used stratify feature to ensure that the training, validation and testing data had the same proportion of samples from each category, based on the category's records proportion to the whole dataset, in order to avoid data skewness to one category, as the data was not balanced and there was a significant difference between the number of samples in classes.

E. DATA AUGMENTATION

The training data, which resulted from the data splitting, was saved in a separate file to apply data augmentation to it as the classes of the dataset were not balanced. For data augmentation, we used ChatGPT by inputting tweets from each class which needed to be augmented and asking for 20 paraphrasing versions for the input text, with the same dialects of the input text. For example, from class "low mood", we selected a tweet which has text that says, "I feel I want to cry", and the ChatGPT gave several outputs like "I feel emotionally distressed" and "I feel so sad and emotionally distress, I want to cry". Following is a sample of the sentences which were generated from this tweet

خذلت نفسي

from the class "feeling worthless":

1. أخفقت في تحقيق هدفي وأشعر بخيبة أمل كبيرة
2. أحس بالخيبة وأنا لم أستطع تحقيق النتيجة المرجوة
3. أشعر بعدم الرضا عن نفسي بسبب الأخطاء التي ارتكبتها

After applying data augmentation for classes 1,2,3,4,5,6 and 7 the dataset became nearly balanced as in Fig. 4.

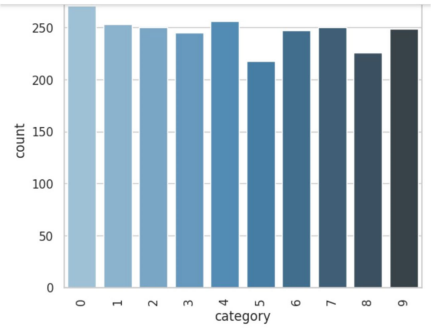

FIGURE 4. Dataset distribution over 10 classes after data augmentation.

After the data augmentation the training data was split to training and validation data. The final number of samples in training dataset is 1628 and the number of samples in validation set is 407.

F. MODEL IMPLEMENTATION AND TRAINING

In this project, we used the pre-trained models AraBERT and MARBERT transformers to apply the classification on ten classes by setting the last layer of the models to output ten probabilities and then labelling the text by taking maximum probability. Before inputting the data into the model, it is important to tokenize the training and validation datasets. In this work, we used the pre-trained AraBERT tokenizer. We set the lowercase parameter to false and split words based on whitespace. The tokenization was implemented on the testing, training and validation datasets, and each record

was returned as a list of input IDs, token type IDs, and attention masks. After that, each value of the corresponding keys (input IDs, token type IDs, and attention masks) was converted to a tensor in addition to each record label which was also converted to a tensor and added to each record in the dictionary. Finally, those tensors were input to the pretrained models. Algorithm 1 summarizes the dataset class which we used to convert training, validation and testing records to tensors, by creating instances of this class for each dataset.

Algorithm 1 Dataset Class

```
class Dataset(torch.utils.data.Dataset)
    initialize(self, encodings, labels=None)
        Set self.encodings to encodings
        Set self.labels to labels
    def getitem(self, idx)
        Initialize an empty dictionary: item
        for each key and val in self.encodings.items()
            Convert val at index idx to a PyTorch tensor
            Add the PyTorch tensor to item with key key
        if self.labels is not None
            Convert self.labels at index idx to a PyTorch tensor
            Add the label tensor to item with key "labels"
        return item
    def len(self)
        return the length of self.encodings["input_ids"]
```

The experiment was run using the hyparameters which are shown in Table 2. The hyperparameters were set based on prior research findings in the literature review. Moreover, we conducted some fine-tuning experiments on the validation set by varying the batch size values, and found that the best performance was as the values set in Table 2. Additionally, we found during the training process, that the loss and accuracy reached stability after epoch 20. Consequently, we set the number of epochs to 20. In Fig. 5, we present the plot for training and validation losses for AraBERT model. The plot shows that the training and validation losses decreased over epochs, indicating that the model was able to learn from the data.

In Fig. 6 we present a concise summary of the methodology steps adopted in this study.

IV. RESULTS EVALUATION

In this section, we will present the results and discussions on comparisons with similar work. In addition, we will present the results of an ablation study we conducted to investigate the effect of data augmentation and adding class 10, which corresponds to the "Normal" class in the dataset.

A. PERFORMANCE METRICS

To evaluate the performance of both architectures, we used the following measuring metrics: accuracy, precision, recall and F1. Where TP means True Positive, TN indicates True Negatives, FP indicates False Positives and FN indicates False Negatives.

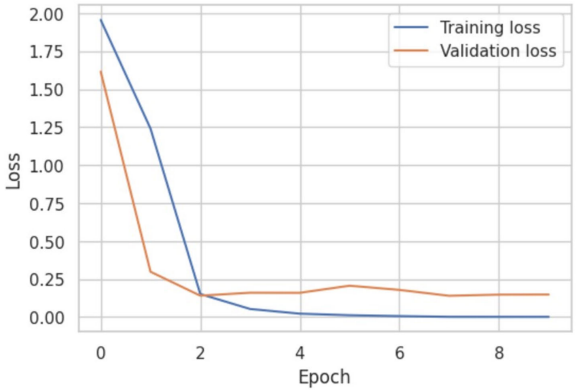

FIGURE 5. Training and validation loses for 20 epochs using AraBERT model.

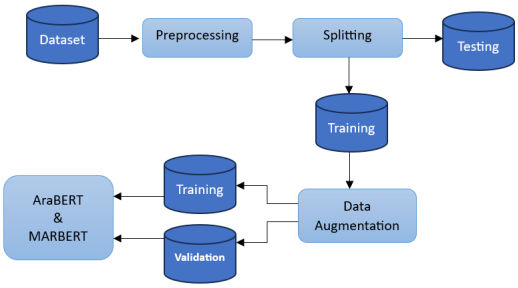

FIGURE 6. Proposed methodolgy: 1) adding normal class to the dataset, 2) data preprocessing 3) split the dataset to training and testing, 4) apply data augmentation on training data, 5) split the training data to training and validation, 6) fine-tune the AraBERT and MARBERT models using training and validation data.

- Accuracy: it indicates the overall true predictions compared to the all predictions which the model has conducted, the value of the accuracy can be achieved using (1):

$$\frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

- Precision: which is the ratio of true positive values to true positive and true negative values. It is an indicator of the model's performance in classifying positive samples. We can calculate the value of precision using (2):

$$\frac{TP}{TP + FP} \tag{2}$$

- Recall: or sensitivity which indicates the ratio of true positive values over the true positive and False Negative.

TABLE 2. Training parameters for AraBERT and MARBERT.

Model	Learning Rate	Batch Size	# of epochs
AraBERT	1e-5	16	20
MARBERT	1e-5	16	20

TABLE 3. Results of AraBERT and MARBERT models.

Model	Accuracy	Precision	Recall	F1
AraBERT	99.3%	99.1%	98.8%	98.9%
MARBERT	98.3%	97.7%	98.2%	98.0%

It measures the model performance in detecting positive samples. We can calculate the value of recall using (3):

$$\frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

- F1: which is harmonic mean that combines the precision and recall scores of the model. It gives a balanced assessment for both precision and recall, based on the literature review we found F1 score is widely used as a performance metric as in [11] and [14]. we can calculate F1 score as in (4):

$$\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

B. RESULTS OF AraBERT AND MARBERT MODELS

The performance of the AraBERT and MARBERT models is shown in Table 3.

From Table 3, we can see that the AraBERT model slightly outperforms the MARBERT model. Nevertheless, Both models achieved promising results in the four performance metrics. In Fig. 7 and Fig. 8, we demonstrate the confusion matrices of MARBERT and AraBERT models. Based on Fig. 7, it can be observed that the MARBERT model successfully classified almost all of the testing data, with the exception of three samples out of a total of 295 samples. These mislabelled tweets belong to classes 2, 3, and 9, which correspond to 'losing interest or pleasure in activities', 'loss of energy', and 'normal class' respectively. Specifically, the first tweet was misclassified as a sleep disorder, the second tweet from the 'loss of energy' class was misclassified as 'low mood', and the final tweet from the 'normal' class was misclassified as 'loss of energy'. On the other hand, there was a slight improvement in the performance of the AraBERT model, which only misclassified one sample out of the 295 testing data samples. This misclassified sample belonged to class 3 which is ''loss of energy'' and was mislabelled as class 4 which is ''low mood''. Nevertheless, to analyze the performance of the models better we printed the mislabeled tweets to examine the specific text that caused confusion in their classification. One of the tweets which was mislabelled is:

ادري والله ملل النوم وملل الجلوس شكلي بتامل السقف لين انام

The tweet can be translated to ''bored of sleeping and sitting, I think I will just stare at the ceiling until I fall asleep''. The tweet is from class 2 which represents ''losing interest or pleasure in activities'', but was mislabelled as class 6 which is ''sleeping disorder''. However, from this tweet, we can recognize that in fact, it can demonstrate the appearance of two symptoms which are sleeping disorder and losing interest, which suggests that a single tweet might reflect the occurrence of two distinct symptoms.

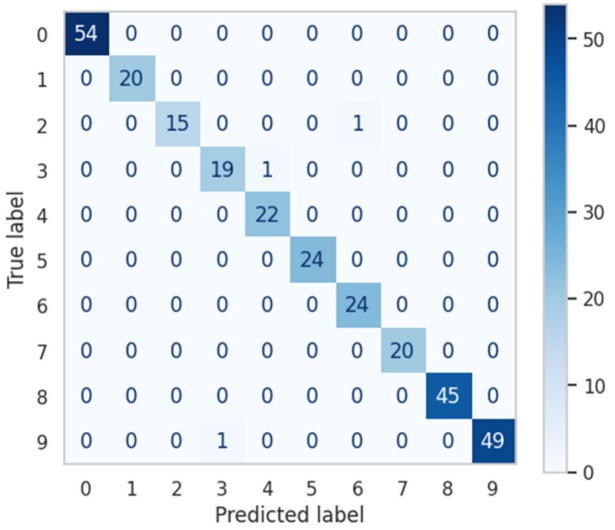

FIGURE 7. Confusion Matrix using the MARBERT Model.

On the other hand, Given the absence of an existing benchmark, this is the first 10-class depression symptom classification approach which utilizes a custom-made dataset. We conducted a comparative analysis with previous studies that utilized AraBERT and MARBERT models for binary classification and 3-class classification tasks, to evaluate the accuracy improvements in classifying different depression symptoms. It is worth mentioning that our dataset was part of the datasets employed by some of these prior works, including the study referenced in [19]. Additionally, other studies, which we compared our work with, such as [11], utilized Arabic tweets for their classification task and utilized the AraBERT and MARBERT models classification task. Nevertheless, we came across an unpublished work that utilized a Multilayer Perceptron (MLP) for 9 depression symptoms classification [27]. Interestingly, this work also employed the same dataset as ours, with the exception that it did not include the 10th class (normal class) that we added. So we compared our proposed model with the model they proposed. On the other hand, depression symptoms detection from the text was also investigated by English language, so we compared our work with [28] which did 9 classes classification to detect depression symptoms, using English Tweets as input text to the proposed model. In Table 4 we summarized the work which we compared our work with, and from Table 4 we can conclude the following notes:

- In comparison to the study conducted by Rabie et al., our utilization of AraBERT and MARBERT models yielded better results than their proposed Multi-Layer Perceptron (MLP) model. Although our task involved 10-class classification, we further experimented by removing the normal class and performed 9-class classification using AraBERT and MARBERT the results are shown in Table 5. Even in this modified

TABLE 4. Comparison between our proposed model and some proposed works in depression and depression symptoms detection.

Proposed Model	Language	Classes	Dataset size	Accuracy	Precision	Recall	F1
Multi-layer Perceptron [27]	Arabic	9	Mood Changing and Depression Dataset	90%	91%	88%	90%
CairoDep(ARABERT) [11].	Arabic	2	7000 Arabic posts	96.93%	96.92%	96.93	96.92
CairoDep(MARBERT) [11].	Arabic	2	7000 Arabic posts	96.07%	96.11%	96.04%	96.07%
AraDepSu(MARBERT) [19].	Arabic	3	more than 10K tweets + modern Arabic mood changing and depression dataset	91.2%	88.74%	88.5%	88.75%
Depression2Vec [28]	English	9	15044 English posts	98%	-	-	98%
our proposed model(AraBERT)	Arabic	10	1223	99.3%	99.1%	98.8%	98.9%
our proposed model(MARBERT)	Arabic	10	1223	98.3%	97.7%	98.2%	98%

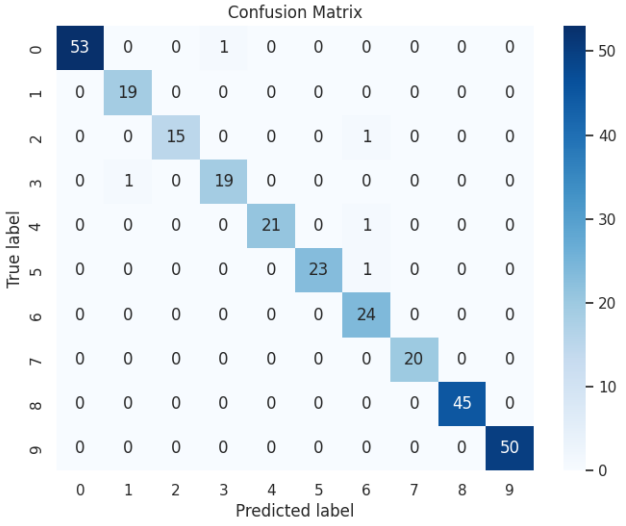

FIGURE 8. Confusion Matrix using the AraBERT Model.

scenario, our models still outperformed the proposed MLP model. These findings provide strong evidence for the effectiveness of pre-trained models over MLP models for this specific task. The improved performance highlights the advantages of leveraging pre-trained models in similar classification tasks.

- Comparing with other work that utilized ARABERT and MARBERT for binary classification and three classes classification we found that AraBERT and MARBERT achieved better results with our custom-made dataset in doing 10-classes classification, Although the proposed work utilized different datasets, it is worth noting that these datasets shared a similarity with ours in terms of containing Arabic tweets. Despite this similarity in dataset nature, our proposed model outperformed the other work, showcasing the efficacy of AraBERT and MARBERT in achieving better results for our 10-class classification task, However, the comparison here is constrained by the differences in the used datasets for each work.
- In comparison to the study conducted by [28] that focused on detecting depression symptoms in the English language, our work presents a different approach. Their proposed model, Depression2Vec, utilized cosine similarity between symptom embeddings from PHQ-9 and patient-authored text. While their results were comparable to ours and demonstrated state-of-the-art performance in depression symptom detection for English text, it is important to acknowledge the limitations of the comparison due to the inherent differences between the English and Arabic languages. These languages possess distinct attributes that can impact the effectiveness and applicability of models and approaches. However, despite the complexities associated with the Arabic language and the limited dataset available, our pre-trained models achieved higher results in detecting the 9 symptoms from Arabic text. This achievement highlights the success of leveraging pre-trained models in overcoming the challenges posed by the Arabic language and demonstrates their effectiveness in addressing depression symptom detection in an Arabic context.

C. ABLATION STUDY

An ablation study was conducted to measure the affect of data augmentation and the affect of adding extra class which is class 10 to the dataset.

Table 5 shows the impact of adding the 10th class and conducting augmentation on the dataset by comparing the accuracy, precision, recall and F1 results before and after applying data augmentation and adding 10th class. The results clearly show the positive impact of adding extra class and data augmentation to the final classification accuracy when using AraBERT model. On the other hand, the results

TABLE 5. Comparsion between models performance after conducting data augmentation and adding class 10.

Model	Accuracy	Precision	Recall	F1
AraBERT	98.46%	98.46%	98.46%	98.46%
AraBERT + class 10	98.72%	98.07%	98.58%	98.25%
AraBERT + augm.+class10	99.2%	99.2%	99.3%	99.2%
MARBERT	96.4%	95.6%	95.0%	95.2%
MARBERT +class 10	98.3%	97.6%	97.9%	97.7%
MARBERT +aug.+class10	98.0%	98.1%	97.9%	98.0%

showed that adding class 10 to the data has increased the results of accuracy, precision, recall and F1 when using MARBERT. However, data augmentation did not make a significant impact on the final results when using AraBERT model

V. CONCLUSION

This work aims to facilitate the process of detecting depression symptoms from social media posts or tweets. In this work, we fine-tuned two pre-trained models: AraBERT and MARBERT, to classify input text into ten classes representing nine depression symptoms and normal cases. The results of this project were promising for further improvements in the awareness and care for those suffering from depression. For future work, It is possible to extend this work by incorporating larger datasets that include an individual's tweet history with timestamps. Furthermore, during the analysis of mislabeled tweets, it was observed that certain symptoms could be associated with multiple labels. Therefore, as a future direction, implementing multi-label classification could be considered. Moreover, emojis can be added to tweets to enhance the accuracy of detecting depression symptoms from Arabic tweets.

Despite achieving promising results, our study has encountered certain limitations. First, the tweets in the used dataset for this project were short and did not contain emojis, as they were deleted by the creators of the dataset, however we believe that emojis and longer text may have a considerable impact on the models' performance. Second, since detecting depression depends on detecting symptoms within a specific period of time, we did not find a dataset which holds history of tweets for specific users(depressed and non depressed users). Finally, to the best of our knowledge, no prior published work exists that directly provides a comparative analysis for the specific research problem or the used dataset in this study.

ACKNOWLEDGMENT

The authors would like to thank the King Fahd University of Petroleum and Minerals for supporting this work.

ETHICS STATEMENT

The dataset used in this study did not include any confidential information about the tweets' authors. The dataset did not show any personal information like their names, their date of births, or any information that may lead to their identities.

The tweets in this dataset may not belong to a depression patients, but still hold one of depression symptoms.

REFERENCES

[1] *COVID-19 Pandemic Triggers 25% Increase in Prevalence of Anxiety and Depression Worldwide*, World Health Org., Geneva, Switzerland, 2022.
[2] M. Zolezzi, M. Alamri, S. Shaar, and D. Rainkie, "Stigma associated with mental illness and its treatment in the Arab culture: A systematic review," *Int. J. Social Psychiatry*, vol. 64, no. 6, pp. 597–609, Sep. 2018.
[3] P. Kaywan, K. Ahmed, A. Ibaida, Y. Miao, and B. Gu, "Early detection of depression using a conversational AI bot: A non-clinical trial," *PLoS ONE*, vol. 18, no. 2, Feb. 2023, Art. no. e0279743.
[4] S. A. S. A. Kulasinghe, A. Jayasinghe, R. M. A. Rathnayaka, P. B. M. M. D. Karunarathne, P. D. S. Silva, and J. A. D. C. A. Jayakodi, "AI based depression and suicide prevention system," in *Proc. Int. Conf. Advancements Comput. (ICAC)*, Dec. 2019, pp. 73–78.
[5] M. M. Aldarwish and H. F. Ahmad, "Predicting depression levels using social media posts," in *Proc. IEEE 13th Int. Symp. Auto. Decentralized Syst. (ISADS)*, Mar. 2017, pp. 277–280.
[6] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial," *JMIR Mental Health*, vol. 4, no. 2, p. e19, Jun. 2017.
[7] F. M. Shah, F. Ahmed, S. K. Saha Joy, S. Ahmed, S. Sadek, R. Shil, and Md. H. Kabir, "Early depression detection from social network using deep learning techniques," in *Proc. IEEE Region 10 Symp. (TENSYMP)*, Jun. 2020, pp. 823–826.
[8] *Instrument: Patient Health Questionnaire (PHQ-9)*, World Health Org., Geneva, Switzerland, 2014.
[9] *Diagnostic and Statistical Manual of Mental Disorders*, American Psychiatric Assoc., Washington, DC, USA, 2013.
[10] A. Maghraby and H. Ali, "Modern standard Arabic mood changing and depression dataset," *Data Brief*, vol. 41, Apr. 2022, Art. no. 107999.
[11] M. El-Ramly, H. Abu-Elyazid, Y. Mo'men, G. Alshaer, N. Adib, K. A. Eldeen, and M. El-Shazly, "CairoDep: Detecting depression in Arabic posts using BERT transformers," in *Proc. 10th Int. Conf. Intell. Comput. Inf. Syst. (ICICIS)*, Dec. 2021, pp. 207–212.
[12] S. H. Aldhafer and M. Yakhlef, "Depression detection in Arabic tweets using deep learning," in *Proc. 6th Int. Conf. Inf. Technol., Inf. Syst. Electr. Eng.*, Dec. 2022, pp. 1–6.
[13] S. Almouzini, M. Khemakhem, and A. Alageel, "Detecting Arabic depressed users from Twitter data," *Proc. Comput. Sci.*, vol. 163, pp. 257–265, Jul. 2019.
[14] 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.
[15] N. Al-Musallam and M. Al-Abdullatif, "Depression detection through identifying depressive Arabic tweets from Saudi Arabia: Machine learning approach," in *Proc. 5th Nat. Conf. Saudi Comput. Colleges (NCCC)*, Dec. 2022, pp. 11–18.
[16] A. Abdulsalam, A. Alhothali, and S. Al-Ghamdi, "Detecting suicidality in Arabic tweets using machine learning and deep learning techniques," *Arabian J. Sci. Eng.*, pp. 1–14, Mar. 2024.
[17] N. A. Baghdadi, A. Malki, H. M. Balaha, Y. AbdulAzeem, M. Badawy, and M. Elhosseini, "An optimized deep learning approach for suicide detection through Arabic tweets," *PeerJ Comput. Sci.*, vol. 8, p. e1070, Jul. 2022.
[18] D. A. Musleh, T. A. Alkhales, R. A. Almakki, S. E. Alnajim, S. K. Almarshad, R. S. Alhasaniah, S. S. Aljameel, and A. A. Almuqhim, "Twitter Arabic sentiment analysis to detect depression using machine learning," *Comput., Mater. Continua*, vol. 71, no. 2, pp. 3463–3477, 2022.
[19] M. Hassib, N. Hossam, J. Sameh, and M. Torki, "AraDepSu: Detecting depression and suicidal ideation in Arabic tweets using transformers," in *Proc. The 7th Arabic Natural Lang. Process. Workshop (WANLP)*, 2022, pp. 302–311.

[20] L. Almuqren and A. Cristea, "AraCust: A Saudi telecom tweets corpus for sentiment analysis," *PeerJ Comput. Sci.*, vol. 7, p. e510, May 2021.
[21] G. Ansari, M. Garg, and C. Saxena, "Data augmentation for mental health classification on social media," 2021, *arXiv:2112.10064*.
[22] V. Tejaswini, K. Sathya Babu, and B. Sahoo, "Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 23, no. 1, pp. 1–20, Jan. 2024.
[23] R. Sawhney, H. Joshi, S. Gandhi, and R. R. Shah, "A time-aware transformer based model for suicide ideation detection on social media," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 7685–7697.
[24] W. Antoun, F. Baly, and H. Hajj, "Arabert: Transformer-based model for Arabic language understanding," in *Proc. Workshop Lang. Resour. Eval. Conf.*, 2020, p. 9.
[25] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT & MARBERT: Deep bidirectional transformers for Arabic," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 7088–7105.
[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.* Minneapolis, MI, USA: Association for Computational Linguistics, vol. 1, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423
[27] E. M. Rabie, A. F. Hashem, and F. K. Alsheref, "Depressive state detection model in Arabic user-generated," doi: 10.21203/rs.3.rs-2281584/v1.
[28] S. K. Mukhiya, U. Ahmed, F. Rabbi, K. I. Pun, and Y. Lamo, "Adaptation of IDPT system based on patient-authored text data using NLP," in *Proc. IEEE 33rd Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jul. 2020, pp. 226–232.

SUZAN ELMAJALI received the B.S. degree in business information systems from the University of Jordan, Jordan, in 2012, and the M.S. degree in computer science from Arizona State University, USA, in 2021. She is currently pursuing the M.S. degree in computer science with the King Fahd University of Petroleum and Minerals, Saudi Arabia. She received her internship at Ernst and Young-Jordan. She worked as a Data Analyst and participated as a fellow, responsible for developing a data analysis system for one of ASU initiatives. Her research interests include machine learning, deep learning, and natural language processing (NLP).

IRFAN AHMAD received the M.S. degree in computer science from KFUPM, Saudi Arabia, in 2008, and the Ph.D. degree in computer science from TU Dortmund, Germany, in 2017. He is currently an Assistant Professor with the Department of Information and Computer Science, KFUPM. He has published several articles in peer-reviewed journal and international conferences. In addition, he has authored a book chapter and three U.S. patents. He has participated in several funded research projects and has been invited to review manuscripts in well-known journals and conferences in his area of expertise. He is also an Academic Editor of *PeerJ Computer Science* journal. His research interests include AI and pattern recognition including machine learning and deep learning, document analysis and recognition, natural language processing, and AI in software engineering.