



Deep Learning for Detecting Mental Disorder from User Content on Social Media

Malek Tounsi¹ · Hanen Ameer¹ · Salma Jamoussi¹

Received: 10 June 2024 / Accepted: 8 January 2025
© The Author(s), under exclusive licence to Springer Nature Singapore Pte Ltd. 2025

Abstract

The prevalence of mental health disorders is rising, with many individuals expressing their thoughts and feelings through social media posts. This presents an opportunity for early detection of mental conditions using natural language processing (NLP). This paper addresses the lack of effective tools for detecting mental disorders from Arabic social media content. We propose **Ara-MentalBERT**, a novel deep learning fusion model that combines AraBERT and MentalBERT using a deep joint autoencoder (DJA). The DJA model aims to integrate Arabic language understanding with specialized mental health domain knowledge by fusing Arabic and mental health embedding representations of text. Our proposed method involves three key steps: (1) fine-tuning pre-trained AraBERT and MentalBERT models, (2) employing a deep joint autoencoder for Ara-Mental text representation learning, and (3) utilizing a mental health disorder classifier with the fused Ara-Mental representations. We evaluate our model on two diverse datasets: a Reddit dataset and a Twitter dataset. Results demonstrate significant improvements over existing methods, achieving an accuracy of 81.12% on the Reddit dataset for combined titles and posts and 86.12% accuracy on the Arabic tweets. These findings highlight the potential of our model to enhance mental health detection in Arabic-speaking communities, offering a powerful tool for identifying at-risk individuals more effectively. Our proposed method makes valuable progress at the intersection of NLP, mental health, and deep learning, paving the way for Arabic social media monitoring to facilitate early intervention and improve mental healthcare.

Keywords Mental health disorders · AraBERT · MentalBERT · Ara-MentalBERT · Ara-Mental representation · Deep joint autoencoder · Deep learning

Introduction

The significance of mental health has become more widely acknowledged globally, as it affects mood, eating habits, social interactions, and contributes to mortality rates. However, a significant barrier remains for many individuals seeking help due to the stigma associated with mental illnesses. Social media platforms like Reddit, Twitter, and Facebook have provided a unique space where people feel

more comfortable discussing their mental health struggles, often anonymously. Recognizing this, the research aims to exploit the vast amounts of user-generated content in Arabic on these platforms to develop a tool that can assist in the early detection of mental health disorders.

One of the primary challenges addressed by this research is the lack of tools and studies focusing on detecting mental health disorders from Arabic social media content. While there is a body of research on this topic in English, the same cannot be said for Arabic, leaving a gap in culturally specific mental health support. This gap not only affects individuals who could benefit from early detection and intervention but also limits the ability of mental health professionals to effectively analyze and screen Arabic social media content for signs of mental disorders.

The primary contribution of this work lies in the development of the Ara-MentalBERT model, specifically dedicated to identifying mental health conditions from Arabic social media posts. The research employs transfer learning

✉ Malek Tounsi
Tounsi.malek@outlook.com

Hanen Ameer
hanene.ameur@isims.usf.tn

Salma Jamoussi
salma.jamoussi@isims.usf.tn

¹ Multimedia Information Systems and Advanced Computing Laboratory, University of Sfax, Technopole of Sfax, Sfax, Tunisia

techniques, building upon the pre-trained AraBERT and MentalBERT models. By fine-tuning these models on a corpus of social media data, the Ara-MentalBERT model aims to enhance the accuracy and effectiveness of detecting a range of mental health disorders. Key novelties of this approach include:

1. Utilizing transfer learning by fine-tuning pre-trained language models, AraBERT and MentalBERT, on a corpus of mental health-related social media data to enhance the detection of mental disorders from Arabic text.
2. Introducing a novel fusion approach that combines the complementary knowledge from AraBERT and MentalBERT using a Deep Joint Autoencoder (DJAЕ). This technique learns a unified representation, termed Ara-MentalBERT, that integrates the linguistic capabilities of AraBERT for Arabic text and the domain-specific knowledge of MentalBERT for mental health-related tasks.

The outcomes of this research hold promise for mental health professionals that can benefit from a tool that aids in the analysis and screening of Arabic social media content, enabling early intervention and support for individuals in need. Moreover, by increasing access to culturally informed mental health support, the research contributes to breaking down the barriers that prevent many from seeking help.

Our work is structured as follows. First, we review related work on detecting mental disorders from social media posts using machine learning and deep learning techniques. Next, we present our proposed Ara-MentalBERT model architecture specialized for Arabic mental health detection. Finally, we describe our experimental setup and evaluation, analyzing the performance of Ara-MentalBERT on identifying specific mental illnesses from user posts.

Related Work

Detecting mental health disorders from social media data has gained significant attention in recent years. Researchers have explored various machine learning (ML) and deep learning (DL) techniques to analyze user-generated content and identify linguistic and behavioral patterns indicative of mental illnesses. This section presents an overview of the state-of-the-art approaches in this domain, covering a broader range of mental health issues.

Machine Learning Based Approach

Traditional machine learning methods, such as Support Vector Machines (SVMs), Naive Bayes, and gradient boosting algorithms, have been applied to social media

data for mental health detection. Kumar et al. [1] proposed an ensemble learning model combining Naive Bayes, Gradient Boosting, and Random Forest classifiers for predicting anxious depression from Twitter data. This approach took advantage of the strengths of multiple algorithms; however, it may have been constrained by its reliance on a single data source and potential overfitting to platform-specific features. In the domain of postpartum depression, Shatte et al. [2] employed SVM classifiers with features describing user behavior, emotions, linguistic styles, and discussion topics to detect postpartum depression from Reddit posts by fathers. This study highlighted the potential of machine learning in identifying subtle indicators of mental health issues, but its focus on a specific demographic may limit its generalizability.

Suicide ideation detection has emerged as a critical area of research in mental health analysis on social media. Chadha et al. [3] conducted a comprehensive study on suicidal ideation from the perspective of social and opinion mining. In a subsequent study, Chadha et al. [4] developed a machine learning-based dataset for finding suicidal ideation on Twitter. They collected and annotated a large corpus of tweets, creating a valuable resource for training and evaluating models designed to detect suicidal thoughts in social media content using eight machine learning algorithms, including Multinomial Naïve Bayes, Bernoulli Naïve Bayes, Logistic Regression, Decision Tree, Support Vector Machine, Random Forest, Voting Ensemble and AdaBoost. This work provided a robust comparison of algorithms but may have been limited by the inherent biases in social media data and the challenges of accurately labeling mental health-related content. In the domain of suicide ideation detection, Mbarek et al. [5] proposed an across online social networks profile building approach. Their study aimed to detect suicidal ideation by constructing user profiles from multiple social media platforms. The authors developed a method to link user accounts across different social networks and aggregate data to create comprehensive user profiles. These profiles were then analyzed using machine learning techniques to identify patterns indicative of suicidal thoughts. However, it also highlights the need for careful consideration of privacy, ethics, and the potential limitations of social media data in mental health assessment.

Nova [6] explored Multinomial Naive Bayes, Multi-layer Perceptron, and LightGBM models for multi-class mental illness classification on Reddit data. This study attempted to address the complexity of mental health classification by considering multiple categories, yet it was potentially constrained by the limitations of traditional machine learning algorithms in capturing the nuanced language used in mental health discussions.

Deep Learning Based Approach

Over the past decade, deep learning has emerged as a significant approach for feature extraction and classification tasks in Artificial Intelligence. It employs neural networks with multiple layers to automatically learn data representations and model complex relationships. This structure enables the models to capture complex hidden patterns within data [7], contrasting with traditional machine learning, which relies on hand-engineered feature extraction pipelines separated from classification algorithms. A key advantage of deep learning is the ability to jointly optimize feature learning and model training in an end-to-end manner.

Deep learning techniques, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have shown promising results in detecting mental health disorders from social media text. Kim et al. [8] developed a CNN-based model combined with XGBoost to analyze Reddit posts and identify users with potential mental disorders. Similarly, Cornn [9] employed a CNN model for classifying Reddit comments as depressive or non-depressive. These studies demonstrated the potential of CNNs in capturing relevant features from social media text for mental health detection.

Expanding on CNN architectures, Aragon et al. [10] proposed a deep attention model with Bag of Sub-Emotions (BoSE) representations for detecting depression and anorexia from text. This approach aimed to incorporate emotional context into the feature representation, potentially improving the model's ability to detect nuanced mental health indicators.

RNNs have also been applied to mental health detection tasks. Bouarara [11] utilized a Recurrent Neural Network (RNN) model for detecting mental health conditions like suicide risk and stress from Twitter data. Building upon this, Shrestha et al. [12] proposed a multi-model RNN-based approach for depression prediction on online user forum datasets, attempting to capture the sequential nature of user posts over time.

Attention mechanisms have become increasingly popular in neural networks, especially for natural language processing (NLP) tasks. These mechanisms take into account long-range dependencies between words and have shown excellent results. Attention mechanisms have been incorporated into neural network architectures to improve mental health detection performance. Almars et al. [13] developed a Bi-LSTM model with attention for detecting depression from Arabic Twitter posts, while Coppersmith et al. [14] combined word embeddings, bidirectional LSTMs, and self-attention layers for identifying signals of suicidal thoughts and behaviors.

The most advanced models of attention-based neural networks are Transformers, which are widely known nowadays.

The advent of transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers) and its variants, has revolutionized mental health disorder detection from social media data. Zeberga et al. [15] proposed a framework utilizing knowledge distillation from BERT for identifying depression and anxiety-related posts.

Albalawi et al. [16] investigated BERT's ability to detect the accuracy of Arabic health-related tweets across a wide range of health-related issues. Despite demonstrating that pretrained transformers can outperform traditional word embeddings, the study is constrained by the use of small expert-labeled datasets and reliance on a single platform, Twitter. Another study conducted by Hu et al. [17] suggested a BERT-based approach for classifying short text content by combining mental features and text vectors. Although the proposed model achieves good performance, it is limited by the small dataset size and lack of external validation.

Farruque et al. [18] discussed the use of BERT with traditional classifiers for detecting depression levels in longer social media posts, showing promise but limited by the small number of depressive tweets available for training.

Researchers have also explored variants of BERT for mental health detection tasks. Putra et al. [19] employed the RoBERTa model for detecting depression levels from Twitter data. As proposed in [20], the authors developed a hybrid model combining LSTM, BERT, and RoBERTa for multi-class mental illness detection on Reddit, achieving high accuracy in identifying individuals with specific mental disorders through Reddit posts analysis, but lack external validation and generalization beyond the platform studied.

Building upon their previous work [3, 4], Chadha et al. [21] proposed also a hybrid deep learning model ACL (attention over convolution and long short-term memory neural network) for effective classification and prediction of suicidal ideation from social network data.

The authors of [22] fine-tuned the MentalBERT model to identify trauma-related issues in Reddit posts, while Grigore et al. [23] developed a MentalBERT model pre-trained on eating disorder data for predicting responses to the Eating Disorder Examination Questionnaire. However, both studies have focused on English social media posts, highlighting the need for more research across other languages and cultural contexts.

As presented in [24], the authors compared MentalBERT to non-contextual models for multi-class mental health classification on Reddit posts. Additionally, the authors of reference [25] demonstrated the effectiveness of the BERTweet model, pretrained on Twitter data, for health-related tweet classification tasks.

The studies discussed above made a significant contribution to examine mental health within social media contexts. However, we find that most of the works have focused on English social media posts. There are limited pre-trained natural

language processing models customized for the Arabic mental healthcare domain, highlighting the need for more research in this area. The diverse range of mental health issues covered in these studies, including depression, anxiety, eating disorders, PTSD, and suicidal ideation, demonstrates the broad applicability of machine learning and deep learning techniques in mental health detection from social media data. Despite these advancements, existing methods have several limitations that justify the need for our proposed model. Most studies focus on English-language content, limiting their applicability to diverse linguistic and cultural contexts, particularly in Arabic-speaking communities. Many models are trained on data from specific social media platforms, potentially limiting their generalizability to other online environments. While deep learning models have improved feature extraction, they may still struggle with understanding the nuanced and context-dependent nature of mental health discussions. The use of social media data for mental health detection raises privacy and consent issues that are not always adequately addressed in existing research. Additionally, many studies rely solely on social media data without corroboration from clinical diagnoses, potentially leading to misclassifications. Mental health datasets often suffer from class imbalance, with fewer examples of positive cases, which can affect model performance. As research in this field continues to advance, there is a clear need for developing more accurate, culturally-sensitive, and ethically sound tools for early detection and intervention in mental health disorders, particularly for Arabic communities. Our proposed Ara-MentalBERT model addresses these limitations and fills critical gaps in existing research. By combining AraBERT, specifically designed for Arabic language processing, with MentalBERT, trained on mental health-related content, our model bridges the gap between Arabic-specific language understanding and mental health domain knowledge. This fusion allows for more accurate detection of mental health issues in Arabic social media content, addressing the lack of culturally sensitive tools for Arabic-speaking communities. Our approach utilizes datasets from multiple social media platforms, including Reddit and Twitter, to train and evaluate the model, improving its ability to generalize across different online environments and reducing platform-specific biases. While our model still relies on social media data, we emphasize the importance of privacy and consent in our research methodology, aiming to develop guidelines for the ethical use of our model in real-world applications.

Proposed Method

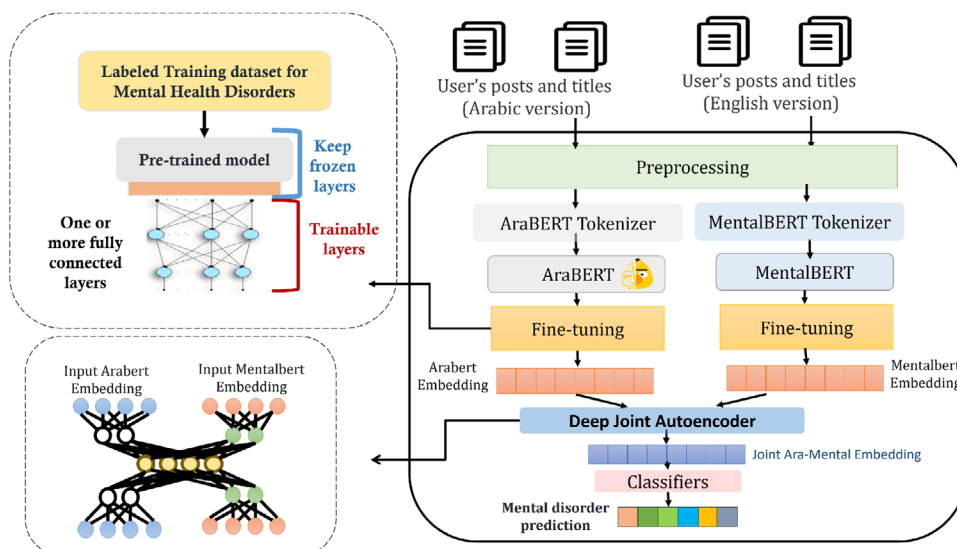
In this section, we introduce a new method for detecting mental disorders from user-generated social media content based on transformers-models, namely AraBERT and MentalBERT. Our proposed method aims to benefit from

the language-specific capabilities of AraBERT for Arabic text and the domain-specific knowledge of MentalBERT for mental health-related tasks. By fusing the representations learned by these models, we aim to create a unified model that can effectively detect mental health disorders from Arabic social media content. To achieve this fusion, we employ a Deep Joint Autoencoder (DJA) architecture. The DJA learns a joint latent representation by reconstructing the input representations from AraBERT and MentalBERT through using two separate autoencoders. This combination can potentially improve the accuracy and robustness of mental disorder detection across both English and Arabic social media posts, addressing the limitations of previous works that focused solely on English data and lacked domain-specific language models for mental healthcare. For our problem statement, we utilize two types of datasets: Reddit dataset (D1) and Arabic tweet dataset (D2).

An overview of our method is depicted in Fig. 1. As shown in Fig. 1, the problem statement is defined as follows: given a collection of social media posts written in English and Arabic $P = \{p_1, p_2, \dots, p_n\}$ and their corresponding titles $T = \{t_1, t_2, \dots, t_n\}$ and their class labels $C = \{c_1, c_2, \dots, c_n\}$, where each class label corresponds to the mental disorder mentioned in titles and posts, the goal is to train a classification model that accurately categorizes new titles and posts. Additionally, for the Arabic tweet dataset (D2), given a collection of tweets $P' = \{p'_1, p'_2, \dots, p'_n\}$ and their class labels $C' = \{c'_1, c'_2, \dots, c'_n\}$. As a first step, we translate an Reddit dataset (D1) to Arabic and the tweet dataset (D2) to English using the Google API. Subsequently, we obtain the English dataset and its Arabic version and process them through the respective tokenizers. Following this step, we prepare two corpus suitable to AraBERT and MentalBERT models. Our proposed method involves three main steps: (i) Fine-tuning of pre-trained models, (ii) Deep joint autoencoder for Ara-Mental text representation learning and (iii) Classification of mental health disorder.

Fine-Tuning of Pre-trained Models

The advent of pre-trained language models has significantly advanced the field of natural language processing. These models provide rich linguistic representations that can be fine-tuned for specific downstream tasks, offering a powerful starting point for various applications. These models, typically based on transformer architectures, learn contextual representations of language from vast corpora of unlabeled text. This pre-training enables the models to capture rich linguistic knowledge and semantic understanding that can be transferred to downstream tasks. One of the most prominent pre-trained models is BERT (Bidirectional Encoder Representations from Transformers), introduced by Devlin et al. [26] in 2018.

Fig. 1 Architecture of the proposed method

Pre-trained Models Employed

In this study, we focus on two pre-trained models based on the BERT (Bidirectional Encoder Representations from Transformers) architecture: AraBERT and MentalBERT.

- **AraBERT** is an Arabic pre-trained language model developed by AUB University, sharing the same configuration as Google's BERT-Base (see Table 1). It comprises 12 encoder blocks, 768 hidden dimensions, 12 attention heads, 512 maximum sequence length, and approximately 136 million parameters [27]. AraBERT was trained on the Open Source International Arabic News Corpus [28], consisting of 70 million Arabic sentences (approximately 24 GB of text). AraBERT excels in comprehending the underlying structure and relationships within the Arabic language by incorporating language-specific pre-processing techniques such as morphological analysis and normalization. Our study focuses on AraBERTv0.2-base, which includes 200 million sentences, 8.6 billion words, and has a model size of 77 GB. This version was trained across 3 million steps using a TPU (Tensor Processing Unit).
- **MentalBERT** is pre-trained on a large corpus of Reddit posts from mental health communities using masked language modeling and next sentence prediction. This provides MentalBERT with a nuanced understanding of language patterns associated with various mental illnesses. Similar to AraBERT, MentalBERT shares the same

architecture and pre-training tasks as BERT. As shown in Table 1, this model follows the standard BERT architecture with 12 attention heads and 768-dimensional hidden states, totaling 110 million parameters [29].

Pre-trained Models Fine-Tuning for Mental Health Disorder Detection Task

Mental health disorder detection is framed as a multi-class text classification task. To adapt pre-trained models for this specific task, we employ fine-tuning techniques. Fine-tuning is a transfer learning method that leverages knowledge gained from one domain to enhance model performance in another. The architecture of our fine-tuned models, as illustrated in Fig. 2, showing how we adapt the pre-trained models (AraBERT and MentalBERT) for our mental health classification task. We conduct extensive hyper-parameter tuning to optimize the fine-tuning process, including adjusting learning rates and batch sizes. Details of fine-tuning are provided in Appendix 6.

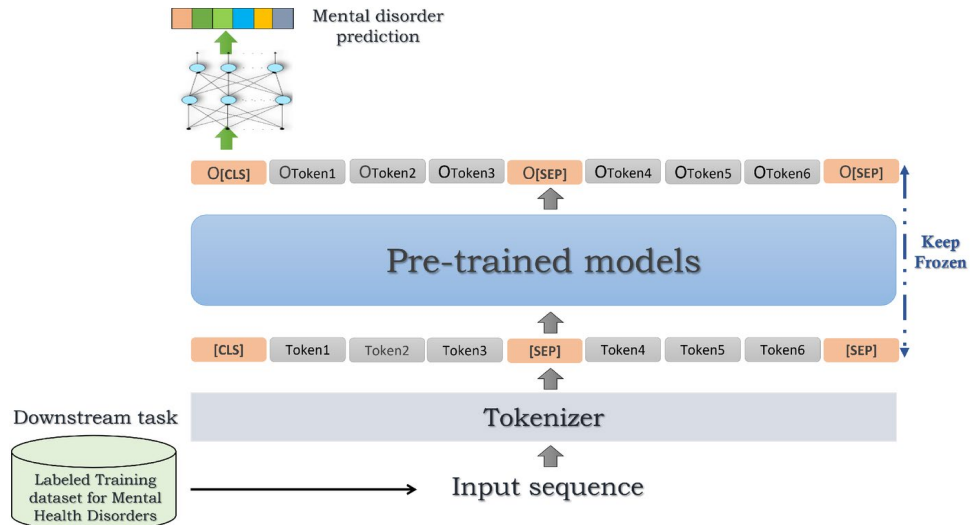
The primary objective of this fine-tuning process is to generate embedding representations that are both representative of Arabic text and sensitive to mental health-related content. Our architecture comprises three key steps:

Table 1 Parameters comparison of pre-trained models

Pre-trained model	L	H	A	Total parameter	Language type
bert-base-arabertv02	12	768	12	136 million	Arabic language
mental-bert-base-uncased	12	768	12	110 million	English language

L numbers of transformer layers, *H* numbers of hidden embedding size, and *A* numbers of attention heads

Fig. 2 Fine-tuning process of pre-trained models for mental disorders prediction



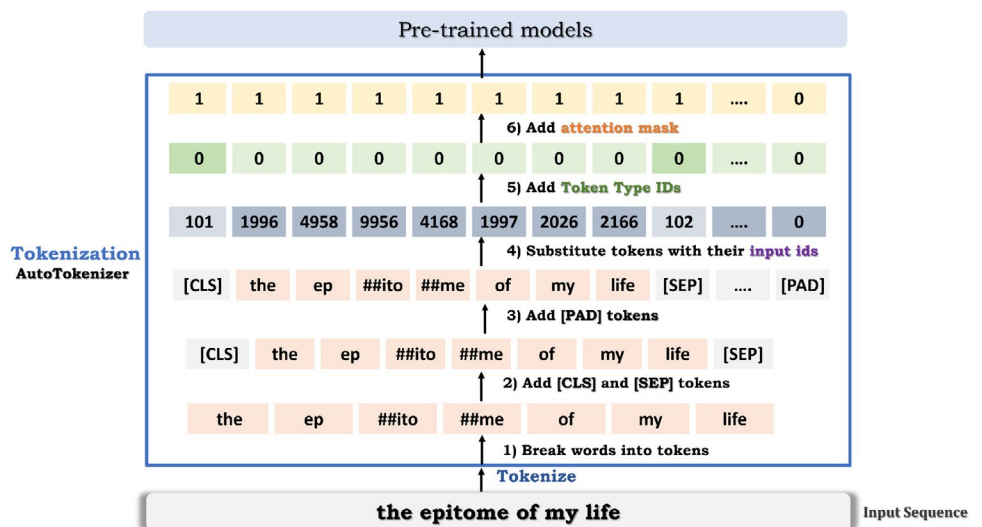
1. **Tokenization:** Data preparation is crucial for effective fine-tuning of pre-trained models. We employ specific tokenizers for AraBERT and MentalBERT to preprocess and tokenize our datasets. Both pre-trained models employ similar tokenization techniques, which is a crucial step in the fine-tuning process. The tokenization procedure includes breaking input text into tokens, adding special tokens ([CLS] at the beginning and [SEP] at the end), converting tokens to their corresponding input IDs, and adding token type IDs and attention masks. This standardized approach allows for consistent processing of text inputs across both models. The tokenization process, illustrated in Fig. 3, involves several steps.
2. **Frozen Pre-trained Layers:** We employ layer freezing during fine-tuning, keeping the weights of the pre-trained transformer layers unchanged. This technique

preserves the general language understanding capabilities acquired during initial pre-training.

3. **Additional Fully Connected Layers:** To adapt the models for mental disorder detection, we add fully connected layers on top of the frozen pre-trained models layers. During training, we utilize two datasets. For AraBERT, we utilize an Arabic dataset of mental health disorders, while for MentalBERT, we use an English dataset of mental health disorders. These datasets are equivalent in content, differing only in language to match the specific pre-trained model. The final [CLS] token representation from the last pre-trained models layers is passed through these additional layers to produce the final classification probabilities.

This fine-tuning approach allows us to create a model that combines the rich linguistic representations learned

Fig. 3 Detailed tokenization process for pre-trained models



by AraBERT and MentalBERT while adapting them to our specific task of mental disorder detection in both Arabic and English texts. By freezing the pre-trained layers and adding task-specific layers, we aim to achieve a balance between retaining general language understanding and learning task-specific features.

Deep Joint Autoencoder for Ara-Mental Text Representation Learning

After fine-tuning AraBERT and MentalBERT on our mental health dataset, our objective is to extract meaningful textual representations, Arabic embedding and mental embedding, learned from Arabert and MentalBERT models to enable efficient classification. The fusion process in our Ara-MentalBERT model is a crucial component that combines the strengths of AraBERT (specialized in Arabic language understanding) and MentalBERT (focused on mental health domain knowledge). In this phase, we leverage the power of context embeddings derived from the pooler output layer of each fine-tuned model and a deep joint autoencoder (DJAЕ) that fuses knowledge from these fine-tuned models to learn a unified text representation.

This fusion is achieved through a novel Deep Joint Autoencoder (DJAЕ) architecture, which learns a unified representation from the heterogeneous embeddings of both models. Figure 4 shows the DJAЕ topology, equivalent to using two separate autoencoders [30], one for AraBERT and one for MentalBERT embeddings, that converge into a shared latent space.

Unlike traditional autoencoders that have a single input, the DJAЕ operates on two input vectors: $a \in \mathbb{R}^d$ representing the AraBERT embedding, and $m \in \mathbb{R}^d$ representing the MentalBERT embedding. These embeddings are extracted from the pooler output layer of each fine-tuned model, capturing rich contextual information about the input text. Each autoencoder's encoder component comprises

multiple layers that progressively reduce the dimensionality of the input embeddings. The final encoded representations from both autoencoders are concatenated in the shared latent space, forming our unified “Ara-Mental” representation. This shared space is designed to capture and integrate the complementary knowledge from AraBERT and mental health domain knowledge in MentalBERT effectively.

Using the joint autoencoder model, we encode each Arabic post text with AraBERT as a continuous vector $v_a \in \mathbb{R}^d$ and each English post text with MentalBERT as $v_m \in \mathbb{R}^d$.

Each autoencoder tries to reconstruct its input by multiple encoding and decoding steps, as shown in the Fig. 4. The encoder part consists of the layers ($L_a^1, L_m^1, L_a^2, L_m^2$, and L_z), while the decoder part consists of layers ($L_{ra}^1, L_{rm}^1, L_{ra}^2, L_{rm}^2$ and L_{rm}^2).

The encoding process involves transforming inputs a and m into lower-dimensional representations h_a and h_m within hidden layers $L_a^1, h_a^1 \in \mathbb{R}^d$ and $h_m^1 \in \mathbb{R}^d$ denote the activation of the hidden layers for a embeddings and m embeddings. Then, a shared hidden layer L_z consists of merging h_a and h_m to produce the joint representation of both embeddings. The fused representation $h_z \in \mathbb{R}^d$ is computed using Eq. (123):

$$h_a^1 = f(W_{ea}^1 \cdot a + b_{ea}^1) \quad (1)$$

$$h_m^1 = f(W_{em}^1 \cdot m + b_{em}^1) \quad (2)$$

$$h_z = f(W_{ea}^2 h_a^1 + W_{em}^2 h_m^1 + b_e^2) \quad (3)$$

where $W_{ea}^{1,2}, W_{em}^{1,2}, b_{ea}^1, b_{em}^1$ and b_e^2 denote the weight matrices and bias vectors.

Next, h_z will be decoded into two disjoint representations h_{ra}^1 and h_{rm}^1 with the same size as the h_a^1 and h_m^1 representations, in order to reconstruct h_a^1 and h_m^1 . Finally, h_{ra}^1 and h_{rm}^1 representations are also decoded into r_a and r_m to reconstruct both mental disorder semantic contextual embedding representations a and m by computing:

$$h_{ra}^1 = f(W_{da}^1 h_z + b_{da}^1) \quad (4)$$

$$h_{rm}^1 = f(W_{dm}^1 h_z + b_{dm}^1) \quad (5)$$

$$r_a = f(W_{da}^2 h_{ra}^1 + b_{da}^2) \quad (6)$$

$$r_m = f(W_{dm}^2 h_{rm}^1 + b_{dm}^2) \quad (7)$$

where $W_{da}^{1,2}, W_{dm}^{1,2}, b_{da}^{1,2}$ and $b_{dm}^{1,2}$ denote the weight matrices and bias vectors. We assume each layer has f as a non-linear activation function (tanh, softmax or relu).

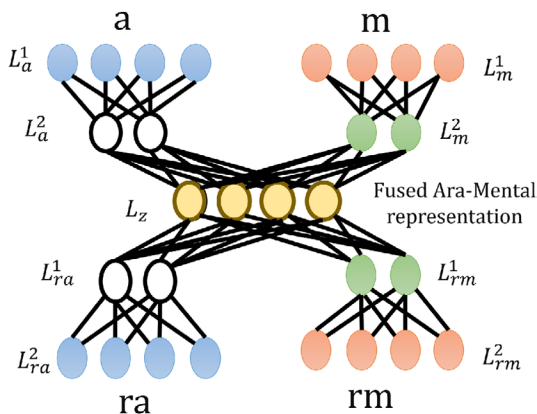


Fig. 4 Ara-Mental text representation using Deep Joint autoencoder

Training the joint autoencoders is achieved by reducing the distance between the original data (input vectors a and m) and its reconstruction (output vectors r_a and r_m). It consists of minimizing the error of reconstructing a^i from r_{a^i} and m^i from r_{m^i} using an equation based on the mean squared error (mse) distance:

$$\text{Err} = \frac{1}{n} \sum_{i=1}^n \left(\|ra^i - a^i\|^2 + \|rm^i - m^i\|^2 \right) \quad (8)$$

Deep joint autoencoder model, adept at handling two inputs and connecting heterogeneous representations, efficiently learns a shared latent space. This adaptability empowers the model to leverage the strengths of both pre-trained language models and demonstrating improved understanding.

To do so, we perform the autoencoder network with two disjoint inputs and outputs for AraBERT and MentalBERT (one perspective each), with separable hidden layers, as illustrated in the Fig. 4. In other words, the two mental health contexts from AraBERT and MentalBERT are available in the input and both are reconstructed. This network includes fully connected hidden layers in common that interact with both Arabic and English mental health contexts to learn a joint representation. The activation of the middle hidden layer is used as a bi-context embedding “fused representation”.

Consequently, the decoder in our DJAE is responsible for reconstructing the two input samples (a) and (m) from the shared latent space (z). This decoding process ensures that the information captured in the joint latent space is utilized to recreate the original contextual representations from AraBERT and MentalBERT.

To further enhance the performance of DJAE, we integrate representations from both AraBERT and MentalBERT within the model’s latent space. This integration fosters a comprehensive understanding of mental health content, as the resulting latent representations encapsulate insights from both language models. This latent representation serve as input to a classifier.

To optimize the Ara-MentalBERT model for mental health analysis, we conducted extensive experiments to adjust the hyperparameters of the deep joint autoencoder as detailed in Appendix 6.3.

Our study leverages deep autoencoders (DJAEs) to detect mental disorders in Arabic populations. DJAEs, known for dimension reduction and improved classification accuracy, bring samples from different classes closer together through dimension reduction. The experiments demonstrate that the optimal configuration of 50 epochs achieves an ideal balance between underfitting and overfitting (see Table 21), maximizing the model’s ability to capture nuanced features related to mental health disorders from Arabic text.

Mental Health Disorder Classification

In the latter case, the pre-trained model undergoes fine-tuning for a classification task, acquiring domain-specific knowledge. Afterwards, it is utilized to represent input text as features. Then, classical machine learning algorithms like SVM, logistic regression and MLP can then utilize these Ara-Mental representation to categorize the posts into mental disorder classes. For mental health disorder classification, we apply two strategies. The first strategy, is using the unified representation from the DJAE as input to SVM, logistic regression and MLP classifiers. The second strategy, utilizes an end-to-end method where the unified representation is connected to additional fully connected layers before prediction. This allows end-to-end training for classification.

Classification Using Joint Ara-Mental Representation

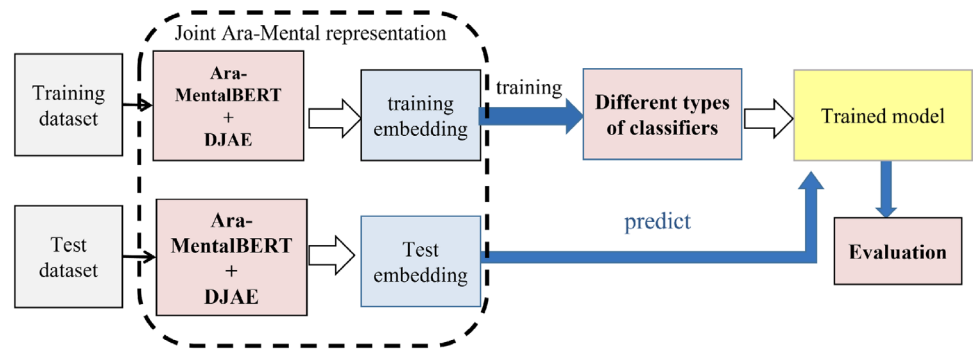
Our classification method leverages the joint Ara-Mental representations learned by the deep joint autoencoder (DJAE) to categorize input texts into predefined mental disorder classes. The DJAE fuses the AraBERT and MentalBERT embeddings into a unified feature vector encapsulating insights from both models. This fused representation encodes the semantic meaning of the text in a joint embedding space. We employ three classifiers, namely Support Vector Machine (SVMs), Logistic Regression (LR) and Multilayer perceptron (MLP) to determine the mental disorder classes for a given text in a supervised manner using its Ara-Mental representation as features. Our MLP takes the fused representations from the joint deep autoencoder like input features as depicted in Fig. 5.

The layers transform these into higher-level abstract representations through non-linear activations. This depth enables progressively extracting relevant properties for the classification task. At the final stage, softmax output layer maps the learned feature embeddings to target mental disorder categories.

By leveraging the unified autoencoder representations, the MLP can categorize text into fine-grained mental health types.

By building on the unified representations, the classifiers can leverage the multimodal knowledge from both Arabic and mental models. This demonstrates the value of joint representation learning through the DJAE for improved mental health classification. The predictions enable fine-grained categorization of texts into specific disorders, allowing targeted community support. To elaborate further, we utilize separate training and test datasets to train the DJAE along with Ara-MentalBERT to obtain joint Aramental representations for both training and testing (see Fig. 5). We then feed the training embeddings to the classifier, which can be SVM, logistic regression, or MLP, to train our model. The

Fig. 5 Classification using Ara-MentalBERT representation with different classifiers



evaluation process is then conducted using the test embeddings to make predictions on the trained model. This classification methodology leverages MLP, SVM or logistic regression on top of the Ara-Mental representations from DJAE to categorize mental health text. The use of separate training and test sets allows us to properly evaluate the model's ability to generalize.

End-to-End Model with Fully Connected Layers

In our second classification strategy, the Deep Joint Autoencoder (DJAE) is directly connected to a classifier, enabling joint optimization of feature learning and prediction. The DJAE encoder portion generates a fused representation h_z of the AraBERT x_a and MentalBERT embeddings x_m using the encoder. This unified representation h_z encapsulating insights from both models is then fed directly into the fully connected layers to predict mental disorder labels y :

$$y = g(W_c h_z + b_c) \quad (9)$$

where g is the classifier activation function, W_c are classifier weights, and b_c is the bias term. Figure 6 shows the end-to-end architecture with the DJAE linked to fully connected classifier layers.

By jointly training the autoencoder and classifier, we facilitate learning a representation optimized for the prediction task. The end-to-end approach enables continuous

refinement of features that contribute to enhanced classification performance. During training, the AraBERT x_a and MentalBERT x_m representations for each training example are fed into the DJAE to generate the unified representation h_z . This latent embedding h_z , called *Ara-Mental*, integrates insights from both models. The Ara-Mental representation is then passed to the fully connected classifier layers for predicting mental disorder labels y .

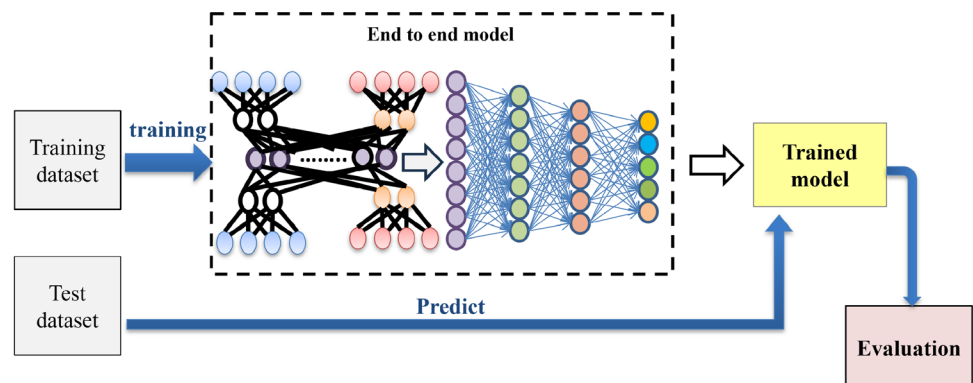
Experimental Results

This section presents the experimental methodology, results, and discussions for our mental health detection models. We begin by describing the datasets used in our experiments. Subsequently, we detail our extensive experiments on fine-tuning AraBERT and MentalBERT models for Arabic and English text classification. Their refined configurations are combined through a deep joint autoencoder to introduce our Ara-MentalBERT model. Additionally, we present the baseline methods used for comparative analysis. Finally, we conclude with a discussion of key findings and results from the experiments.

Datasets

To validate the proposed model, we utilized two datasets: one in English (D1) and one in Arabic (D2).

Fig. 6 End-to-end model with fully connected layers



- Dataset D1, published by Murarka et al. [20], consists of Reddit posts categorized into six classes: ADHD, anxiety, bipolar disorder, depression, PTSD, and a “none” category for unrelated posts. The dataset is pre-divided into training (13,727 posts), validation (1488 posts), and testing (1488 posts) sets.
- Dataset D2, the Arabic Tweets Depression Dataset published by Maghraby et al. [31], comprises Modern Standard Arabic tweets for estimating depression risk in online personas. It contains 976 tweets for training and 245 for testing, categorized into nine depression-related symptoms.

Both datasets provide a diverse range of mental health-related content, allowing for comprehensive evaluation of our proposed Ara-MentalBERT model across different languages and mental health categories. All the details of the number of samples in the training set, validation set and testing set can be seen in Table 2. Tables 3 and 4 offer additional statistics for datasets D1 and D2, respectively.

Dataset D1 exhibits variation in average word count, sentence count, and text length across different subreddits, as shown in Table 3.

These metrics provide a quantitative basis for understanding how individuals express themselves when discussing mental health issues on social media platforms.

- **Average Word Count:** This metric indicates the typical number of words used in titles and posts. It is calculated by summing the word counts of all samples (titles or posts) in a group (like a class) and dividing by the total number of samples. For instance, r/None has the highest average word count for both titles (15.75) and posts (239.43), suggesting more verbose discussions. In contrast, r/Depression has the lowest average word count for posts (152.62), which might reflect the succinct nature of expressions in this community.
- **Average Sentence Count:** This measure reflects the complexity of language structure. It is calculated by summing up the sentence counts of all text samples in a given group and then dividing by the total number of

Table 2 Datasets description

Datasets	Class	Training	Validation	Testing
D1 [20]	Adhd	2465	248	248
	Anxiety	2422	248	248
	Bipolar	2407	248	248
	Depression	2450	248	248
	Ptsd	2001	248	248
	None	1982	248	248
	Total	13,727	1488	1488
D2 [31]	Diminished ability to think or concentrate (DATC)	212	–	59
	Weight disorder (WD)	181	–	45
	Psychomotor agitation or retardation (PAR)	97	–	23
	Sleep disorder (SD)	93	–	24
	Low mood (LM)	87	–	22
	Loss of energy (LE)	84	–	17
	Suicidality (SU)	80	–	20
	Feelings of worthlessness (FW)	82	–	15
	Losing interest or pleasure in activities (LIPA)	60	–	20
	Total	976	–	245

Table 3 D1 dataset: statistics

Subreddit	Samples	Avg word count		Avg sentence count		Avg text length	
		Titles	Posts	Titles	Posts	Titles	Posts
r/None	2478	15.75	239.43	1.52	14.19	94.25	1392.23
r/ADHD	2961	13.75	199.10	1.32	12.69	75.04	1070.81
r/Depression	2946	12.19	152.62	1.22	10.43	65.14	793.31
r/PTSD	2497	10.12	233.95	1.22	14.92	56.28	1233.27
r/Anxiety	2918	11.76	171.40	1.19	10.91	63.95	901.43
r/Bipolar	2903	9.28	203.56	1.22	13.63	51.53	1083.70

Table 4 D2 dataset: statistics

Category	Avg word count	Total words	Tweet count	Avg sentence count	Total sentences	Avg tweet length	Max tweet length
DATC	8.61	2333	271	1.00	271	44.93	181
FW	6.63	650	98	0.99	97	31.43	88
LIPA	4.44	355	80	1.00	80	20.30	96
LE	6.43	649	101	1.00	101	32.41	106
LM	5.91	644	109	1.00	109	29.02	96
PAR	7.62	914	120	1.00	120	38.22	125
SD	7.20	842	117	1.00	117	35.87	88
SU	5.18	518	100	1.00	100	25.21	68
WD	9.43	2132	226	1.00	226	45.13	140

samples in that group. *r*/None shows the highest average sentence count for titles (1.52), indicating more complex title structures. For posts, *r*/PTSD has the highest average (14.92), suggesting more detailed or narrative-driven content.

- **Average Text Length:** This metric represents the average number of characters in a text sample, including spaces and punctuation. It is calculated by summing up the character counts of all text samples in a given group and then dividing by the total number of samples in that group. *r*/None has the highest average text length for both titles (94.25) and posts (1392.23), aligning with its higher word counts. Conversely, *r*/Bipolar has the lowest average title length (51.53), potentially indicating more concise post titles in this community.

These variations across subreddits may reflect differences in communication styles, content complexity, or the nature of discussions within each mental health community. For instance, the longer posts in *r*/PTSD might indicate a tendency for users to share detailed personal experiences, while the shorter posts in *r*/Depression could suggest more concise expressions of emotional states. These linguistic patterns could inform the development of more nuanced natural language processing models for mental health analysis on social media platforms.

Dataset D2 exhibits diverse tweet characteristics across different symptom categories, as illustrated in Table 4.

This table provides valuable insights into the linguistic patterns associated with various mental health symptoms expressed on Twitter. The table presents several key metrics for each symptom category. The average word count offers a measure of tweet verbosity, indicating how elaborately users tend to express themselves when discussing specific symptoms. For instance, the Weight Disorder (WD) category shows the highest average word count (9.43), suggesting that

users often provide more detailed descriptions when discussing weight-related issues. Conversely, the Losing Interest or Pleasure in Activities (LIPA) category has the lowest average word count (4.44), indicating more concise expressions for this symptom. The average sentence count, consistently close to 1.00 across most categories, suggests that users typically express their thoughts in a single sentence per tweet, regardless of the symptom discussed. This pattern may be influenced by Twitter's character limit, encouraging concise communication. Average tweet length, measured in characters, combines aspects of word count and word length, providing an overall measure of content volume. The Weight Disorder and Diminished Ability to Think or Concentrate (DATC) categories show the highest average tweet lengths (45.13 and 44.93 characters, respectively), indicating more verbose expressions for these symptoms. The maximum tweet length varies considerably across categories, ranging from 68 characters for Suicidality (SU) to 181 for DATC. This variation highlights the diversity in expression styles and the potential for more elaborate descriptions within certain symptom categories. These linguistic patterns offer crucial insights for developing more nuanced natural language processing models in mental health analysis. The varying complexity and length of expressions across symptom categories can inform the design of more accurate text classification algorithms and symptom detection tools. Moreover, these findings could guide the development of user interfaces for mental health apps, helping to set appropriate length limits for symptom reporting forms and fine-tune language models to expect certain text lengths for different symptom categories. Understanding these linguistic nuances is essential for creating more effective and sensitive tools for detecting and addressing mental health issues through social media analysis. By recognizing the unique expression patterns associated with different symptoms, researchers and developers can enhance the accuracy and efficacy of automated mental health assessments on social media

platforms, ultimately contributing to improved early detection and intervention strategies in digital mental health care.

Fine-Tuning of Pre-trained Models Assessment

In this section, we provide detailed results of fine-tuning of two pre-trained models, AraBERT and MentalBERT, on the task of mental health disorder classification.

Fine-Tuning AraBERT Results

After AraBERT hyperparameters tuning as detailed in the Appendix 6, the optimized AraBERT model (see “[AraBERT Hyperparameters Adjustment](#)”) is evaluated on the Reddit test dataset (D1) [20] using three different inputs: titles, posts and combination of titles + posts, and on the Twitter test dataset (D2) [31] using tweets only.

The results of our model indicate that accuracy and recall are identical. However, recognizing that accuracy alone is not enough to effectively compare the performance of different algorithms, the authors followed the approach used by Chadha et al. [21] to evaluate model performance. They calculated precision, recall, specificity, and F1 score, providing a more comprehensive understanding of the model’s strengths and weaknesses across different mental health categories and input types, making these metrics a more reliable standard for performance comparison. Analyzing the class-wise as depicted in Tables 5 and 6 highlights the performance of text classification models.

Table 5 shows the importance of complete textual content (posts with their titles) providing more context for accurate classification. Precision, recall, and F1-score are higher for most classes when using combined titles + posts compared to each alone input. However, performance significantly decrease when only titles or posts are utilized, showing the challenge of limited text input. The average F1-score across all classes is 0.74 for the combined titles and posts, significantly higher than using titles (0.61) or posts (0.70) alone.

Interestingly, the specificity (S) values are consistently high across all input types, with the combined titles + posts achieving the highest average specificity of 0.94. This

indicates the model’s strong ability to correctly identify negative cases (i.e., instances not belonging to a particular class). The high specificity, especially for classes like bipolar (0.99) and anxiety (0.96) with combined input, suggests that the model is particularly adept at avoiding false positives for these conditions.

The confusion matrix is a valuable tool for evaluating model performance, especially in classification tasks. It provides deeper insights into the model’s accuracy, errors, and weaknesses.

Figure 7 shows the confusion matrices for the AraBERT model’s performance on the Reddit dataset (D1) using three different inputs: (a) titles, (b) posts, and (c) posts + titles. The combined input of posts and titles provides the most comprehensive information, leading to the best classification performance. This underscores the importance of context in mental health classification tasks.

The model consistently performs well in identifying bipolar disorder across all input types. This suggests that bipolar disorder may have more distinct linguistic markers compared to other conditions in this dataset. The progression from titles to posts to posts+titles demonstrates the value of having complete textual data for accurate mental health classification. This has implications for data collection and model deployment in real-world scenarios.

The results presented in Table 6 and Fig. 8 showcase the impressive performance of the AraBERT model on the Arabic Tweets Depression Dataset (D2).

The Table 6 reveals consistently high precision, recall, F1-scores and specificity across all categories, with macro-averaged scores of 0.94 for precision, 0.92 for recall, 0.93 for F1-score, and an outstanding 0.99 for specificity, indicating excellent overall performance in this multi-class classification task.

Particularly noteworthy is the perfect performance ($P = 1.00$, $R = 1.00$, $F1 = 1.00$) achieved for the “FW” (Feelings of Worthlessness) category, suggesting the model’s exceptional ability to identify this specific class. Other categories such as “DATC” (Diminished Ability to Think or Concentrate), “LE” (Loss of Energy), and “WD” (Weight Disorder) also demonstrate remarkably high F1-scores above 0.95,

Table 5 Results: AraBERT Class-wise precision, recall, F1-score, and specificity for the Reddit dataset (D1) [20] translated on Arabic classification

Class	Titles				Posts				Posts + titles			
	P	R	F1	S	P	R	F1	S	P	R	F1	S
ADHD	0.48	0.56	0.51	0.83	0.52	0.69	0.59	0.85	0.63	0.76	0.69	0.89
anxiety	0.71	0.57	0.63	0.93	0.80	0.68	0.73	0.95	0.79	0.67	0.72	0.96
bipolar	0.85	0.84	0.84	0.95	0.93	0.94	0.94	0.98	0.96	0.92	0.94	0.99
depression	0.54	0.49	0.51	0.88	0.69	0.53	0.60	0.94	0.66	0.61	0.64	0.93
ptsd	0.52	0.60	0.56	0.85	0.59	0.69	0.63	0.88	0.68	0.77	0.72	0.91
none	0.57	0.56	0.57	0.88	0.76	0.68	0.72	0.94	0.77	0.72	0.74	0.95
Average	0.61	0.60	0.61	0.89	0.72	0.70	0.70	0.92	0.75	0.74	0.74	0.94

Table 6 Results: AraBERT Class-wise precision, recall, F1-score and specificity for the Arabic Tweets Depression Dataset (D2) [31] classification

Category	P	R	F1	S
DATC	0.96	0.96	0.96	0.99
FW	1.00	1.00	1.00	1.00
LE	1.00	0.89	0.94	0.99
LIPA	0.94	0.85	0.89	0.99
LM	0.90	0.86	0.88	0.99
PAR	1.00	0.79	0.88	0.98
SD	0.70	1.00	0.83	0.96
SU	0.93	1.00	0.97	0.99
WD	0.97	0.97	0.97	0.99
Macro avg.	0.94	0.92	0.93	0.99

showcasing the model's strength in accurately classifying these categories.

The high specificity scores across all categories (ranging from 0.96 to 1.00) indicate the model's exceptional ability to

correctly identify negative cases. This is particularly important in a clinical context, as it suggests a low false positive rate, which is crucial for avoiding unnecessary interventions or misdiagnoses.

While most categories show balanced precision and recall, some variations are observed. For instance, the “SD” (Sleep Disorder) category shows perfect recall (1.00) but lower precision (0.70), suggesting the model may occasionally overpredict this category. Conversely, “PAR” (Psychomotor Agitation or Retardation) shows perfect precision but lower recall (0.79), indicating some instances of this category may be missed.

The confusion matrix in Fig. 8 provides a more detailed view of the model's classification performance.

It reveals that the model excels at correctly identifying most categories, as evidenced by the high values along the diagonal. The “Diminished ability to think or concentrate” category shows the highest number of correct classifications, with 53 instances accurately identified. Some minor misclassifications are observed, such as a few instances of “losing interest or pleasure in activities” being incorrectly classified

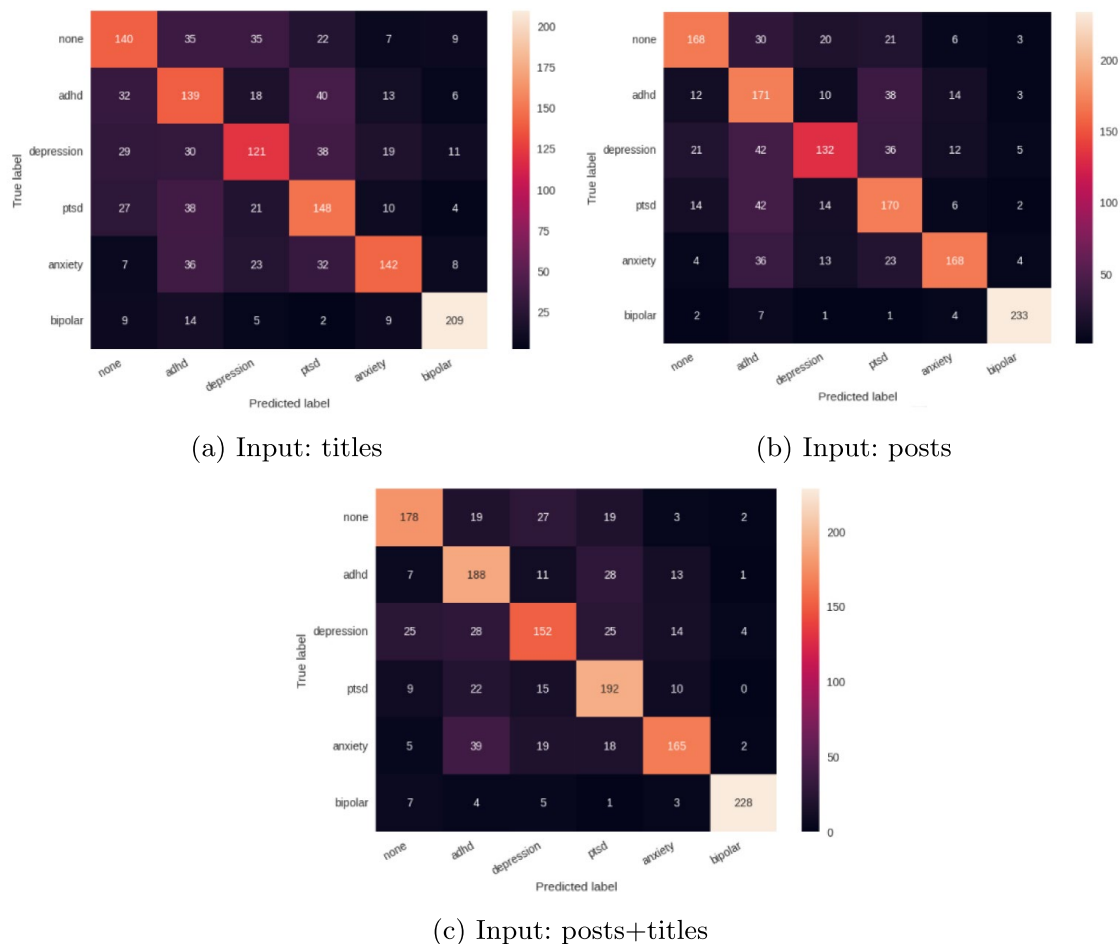
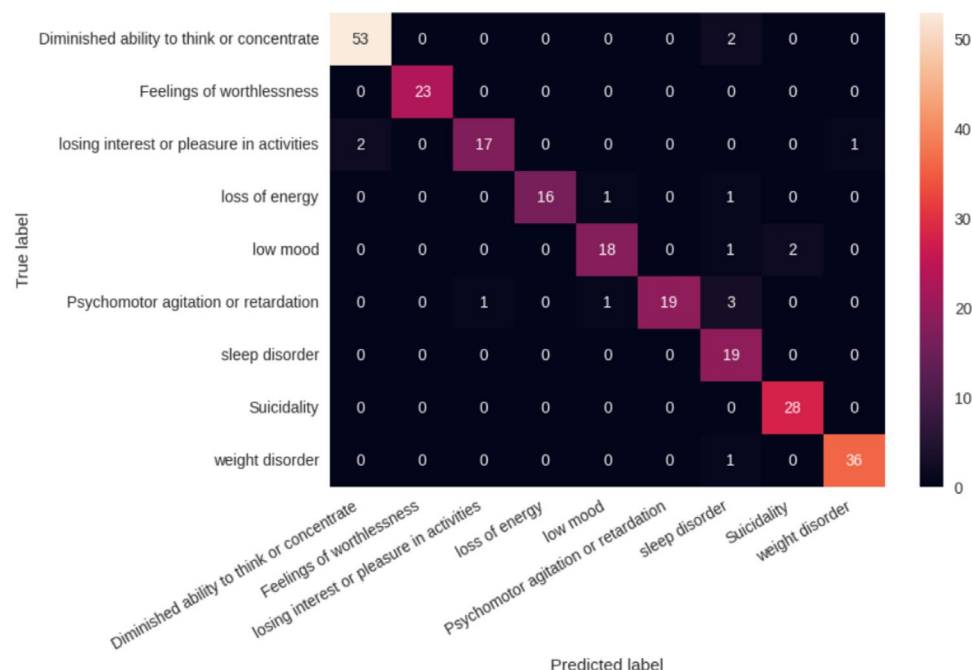
**Fig. 7** AraBERT: Confusion Matrices of Reddit dataset (D1)

Fig. 8 AraBERT: Confusion Matrice of the Arabic Tweets Depression Dataset (D2)



as “Diminished ability to think or concentrate” or “weight disorder”. Similarly, there are occasional misclassifications between “low mood”, “Psychomotor agitation or retardation”, and “sleep disorder”. However, these misclassifications are minimal and do not significantly impact the overall high performance of the model.

The model’s ability to accurately distinguish between closely related symptoms of depression is particularly impressive. For instance, it shows a strong capability to differentiate between “low mood” and “feelings of worthlessness”, which could potentially have similar linguistic expressions.

The high performance across all categories suggests that the AraBERT model has successfully captured the nuances of Arabic language expressions related to various aspects of depression. This is especially noteworthy given the complexity of mental health-related language and the potential cultural and linguistic specifics of expressing such concepts in Arabic.

In Table 6, the classification performance is generally higher compared to the Reddit dataset results in Table 5. Additionally, the overall macro-averaged F1-score is 0.93, which is excellent for a multi-class classification task.

These results indicate that the AraBERT model, when applied to the Arabic Tweets Depression Dataset, demonstrates robust and reliable performance in identifying various depression-related categories. Such a tool could prove invaluable in mental health research, early detection systems, and potentially in supporting clinical assessments in Arabic-speaking contexts. However, it’s important to note that while the model shows excellent performance, any application in

clinical settings should be done cautiously and in conjunction with professional medical assessment.

Fine-Tuning MentalBERT Results

After several steps of tuning the MLP implementation, adjusting the number of epochs and the batch size as mentioned in “[MentalBERT Hyperparameters Adjustment](#)” on Appendix 6, the final Architecture 3 of MentalBERT model ($768 > 400 > 200 > 100 > 6$) is evaluated on the test Reddit dataset (D1) [20] across three data inputs: Titles, Posts and Posts + Titles and on the test Tweet dataset (D2) [31].

The results presented in Table 7 provide valuable insights into the performance of the MentalBERT model in classifying mental health-related posts on Reddit dataset (D1). The table illustrates the model’s performance across three input configurations: Titles, Posts, and Posts + Titles, with metrics including precision (P), recall (R), F1-score (F1), and specificity (S).

Some key observations from the Table 7 demonstrates performance on most categories is higher when using both titles and posts compared to either source alone. This highlights the value of more input text for MentalBERT’s contextual embeddings. This is evident in the highest average precision (0.82), recall (0.81), F1-score (0.81), and specificity (0.97) for the Posts + Titles configuration. This underscores the value of comprehensive input for MentalBERT’s contextual embeddings.

The addition of specificity metrics shows the model’s strong ability to correctly identify negative cases across all categories and configurations. The average specificity is

Table 7 Results: MentalBERT Class-wise precision, recall, F1-score, and specificity for Reddit dataset (D1) [20] classification

Class	Titles				Posts				Posts + titles			
	P	R	F1	S	P	R	F1	S	P	R	F1	S
ADHD	0.67	0.60	0.64	0.94	0.67	0.72	0.77	0.95	0.75	0.78	0.77	0.96
Anxiety	0.74	0.51	0.61	0.96	0.77	0.75	0.76	0.97	0.85	0.73	0.79	0.98
Bipolar	0.88	0.91	0.90	0.99	0.95	0.94	0.95	0.99	0.96	0.98	0.97	0.99
Depression	0.56	0.60	0.58	0.92	0.73	0.65	0.69	0.95	0.81	0.71	0.76	0.97
Ptsd	0.57	0.79	0.66	0.93	0.64	0.76	0.70	0.94	0.69	0.85	0.76	0.95
None	0.72	0.66	0.69	0.95	0.82	0.72	0.77	0.97	0.85	0.81	0.83	0.98
Average	0.69	0.68	0.68	0.95	0.76	0.76	0.76	0.96	0.82	0.81	0.81	0.97

notably high, ranging from 0.95 for Titles only to 0.97 for Posts + Titles. This indicates a low false positive rate, which is crucial in mental health applications to avoid unnecessary interventions or misdiagnoses. The “bipolar” category consistently shows the highest performance across all metrics and configurations, with near-perfect scores in the Posts + Titles setup (P: 0.96, R: 0.98, F1: 0.97, S: 0.99). This suggests the model’s particular strength in identifying bipolar disorder-related content.

The “depression” and “ptsd” categories show lower performance, especially in the Titles-only configuration. This might indicate that these conditions require more context (provided by full posts) for accurate classification.

In the Posts + Titles configuration, most categories show well-balanced precision and recall, leading to strong F1-scores. This balance is crucial for practical applications, ensuring a good trade-off between false positives and false negatives.

While specificity is consistently high across all categories and configurations, other metrics show more variation. This suggests that while the model is generally good at identifying negative cases, there’s room for improvement in correctly identifying positive cases for some categories.

Samely to AraBERT model, progression from Titles-only (accuracy: 68.01%, F1: 67.83%) to Posts-only (accuracy: 75.74%) to Posts + Titles demonstrates the advantage of utilizing complete text for improving classification performance. The high specificity across all configurations further solidifies the model’s reliability in minimizing false positives. These results highlight MentalBERT’s potential for accurate mental health content classification, particularly when provided with comprehensive input (posts and titles). The model’s strong performance, especially in specificity, suggests its potential utility in automated screening or monitoring applications for mental health conditions, though clinical validation would be necessary for any practical implementation.

Based on the confusion matrices presented in Fig. 9, MentalBERT’s performance varies across different input types. The model’s performance improves from using titles only, posts only, and combined titles and posts, suggesting

that more comprehensive input (likely combined titles and posts) leads to better classification accuracy. Across all input types, MentalBERT excels at identifying bipolar disorder. This suggests that the linguistic markers for bipolar disorder may be more distinct than those of other conditions. The confusion between different mental health conditions, particularly when using limited input (titles only), highlights the need for cautious interpretation of automated classification results in clinical settings.

The results presented in Table 8 and the accompanying confusion matrix in Fig. 8 provide valuable insights into the performance of MentalBERT on the English-translated tweet dataset (D2) for depression classification.

MentalBERT demonstrates strong overall performance across the various categories of depression symptoms, with a macro-averaged F1-score of 0.85 and an impressive macro-averaged specificity of 0.99. This indicates that the model is generally effective at distinguishing between different aspects of depressive symptomatology in social media text and excels at correctly identifying negative cases.

The model excels particularly in identifying weight disorder (WD), with perfect recall and specificity, and near-perfect precision (0.97, 1.00, 0.99, and 1.00 respectively for P, R, F1, and S). This suggests that language related to weight issues in the context of depression is highly distinctive and easily recognizable by the model, with no false positives.

Suicidality (SU) is another category where MentalBERT performs exceptionally well, with an F1-score of 0.93 and perfect specificity (1.00). This is a crucial finding, given the critical nature of identifying suicidal ideation in mental health monitoring. The perfect specificity indicates that the model is highly reliable in not misclassifying non-suicidal content as suicidal. The model also shows strong performance in recognizing diminished ability to think or concentrate (DATC), with an F1-score of 0.89 and specificity of 0.99. This is reflected in the confusion matrix, where 50 out of 55 instances are correctly classified, with minimal misclassifications spread across other categories. Categories such as feelings of worthlessness (FW), loss of energy (LE), and sleep disorder (SD) all show solid performance with F1-scores above 0.80 and specificity of 0.99. This indicates

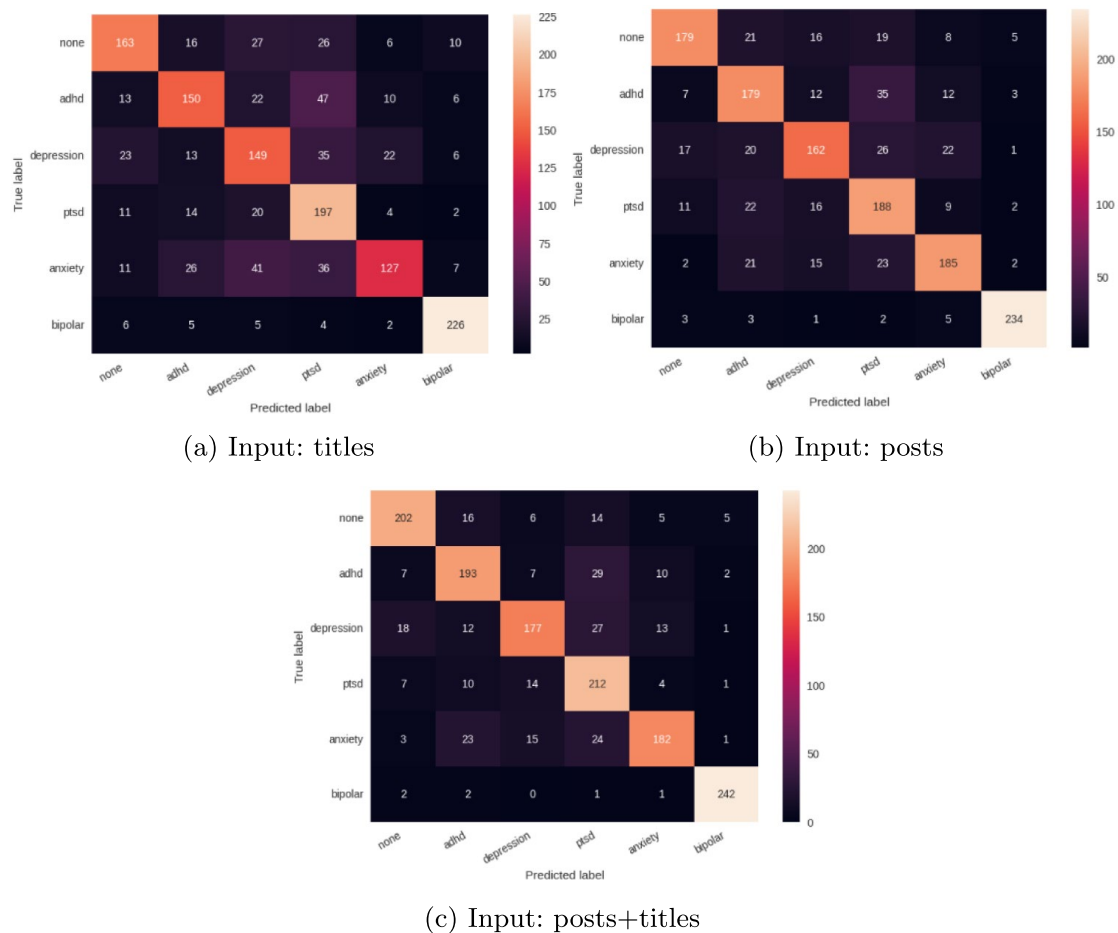


Fig. 9 MentalBERT: Confusion Matrices of Reddit dataset (D1)

Table 8 Results: MentalBERT Class-wise precision, recall, F1-score, and specificity for the tweet dataset (D2) [31] translated on English classification

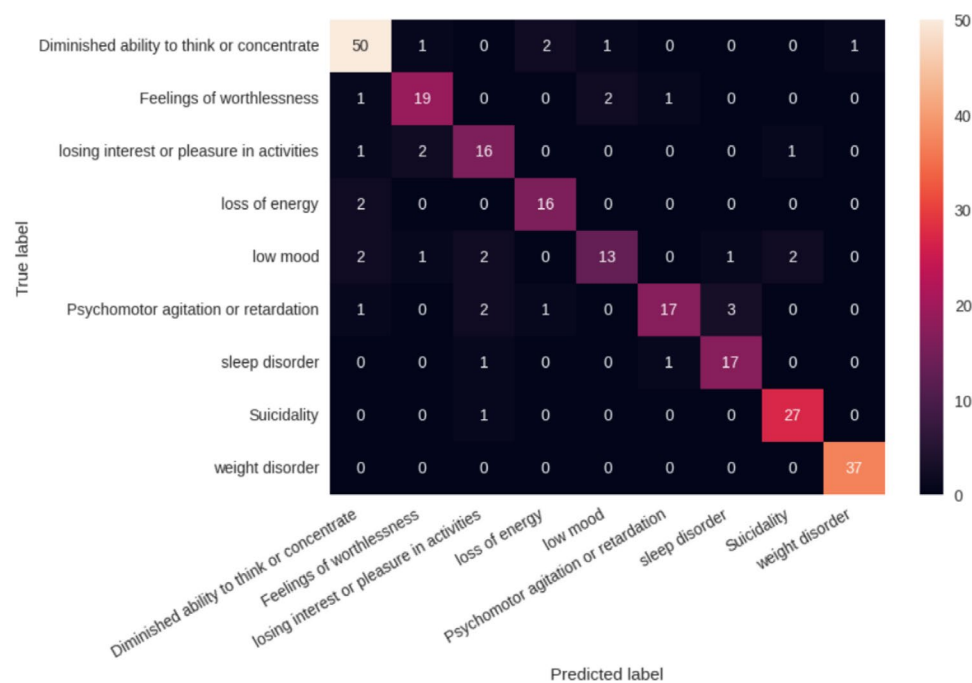
Category	P	R	F1	S
DATC	0.88	0.91	0.89	0.99
FW	0.83	0.83	0.83	0.99
LE	0.84	0.89	0.86	0.99
LIPA	0.73	0.80	0.76	0.98
LM	0.81	0.62	0.70	0.99
PAR	0.89	0.71	0.79	0.99
SD	0.81	0.89	0.85	0.99
SU	0.90	0.96	0.93	1.00
WD	0.97	1.00	0.99	1.00
Macro avg.	0.85	0.85	0.85	0.99

that MentalBERT can reliably identify these distinct symptoms of depression from textual data while maintaining a very low false positive rate.

The lowest performing category is low mood (LM) with an F1-score of 0.70, primarily due to lower recall (0.62). However, its high specificity (0.99) suggests that when the model does identify low mood, it is highly likely to be correct. The consistently high specificity across all categories (ranging from 0.98 to 1.00) is particularly noteworthy.

The confusion matrix (see Fig. 10) reveals some challenges in distinguishing between closely related symptoms. For instance, there is some confusion between losing interest or pleasure in activities (LIPA) and other categories like low mood (LM) and psychomotor agitation or retardation (PAR). This is not unexpected, given the potential overlap in linguistic expressions of these symptoms. Low mood (LM) shows the lowest F1-score at 0.70, with a notably lower recall (0.62) compared to its precision (0.81). The confusion matrix indicates that low mood is occasionally misclassified as other categories, particularly DATC and LIPA. This suggests that expressions of low mood may share linguistic features with other depressive symptoms, making it more challenging for the model to distinguish.

Fig. 10 MentalBERT: Confusion Matrice of Arabic Tweets Depression Dataset (D2)



Despite these challenges, MentalBERT's performance is impressive, especially considering the complexity of mental health language and the potential noise introduced by translation from Arabic to English. The model's ability to maintain high performance across most categories indicates its robustness in capturing depressive symptoms from social media text.

These results suggest that MentalBERT could be a valuable tool for large-scale screening and monitoring of depressive symptoms in social media contexts. However, it's important to note that while the model shows strong performance, any application in clinical settings should be done cautiously and in conjunction with professional medical assessment. Future work might focus on improving the model's ability to distinguish between closely related symptoms, particularly in the areas where some confusion was observed.

Overall, the MentalBERT model designed through Architecture 3 and hyperparameter tuning is able to achieve strong mental health detection, especially when provided with sufficient text from both posts and titles. The results demonstrate the importance of input length for MentalBERT.

Mental Health Disorder Classification Assessment

The optimized joint autoencoder model is evaluated for Ara-Mental health classification through two strategies:

1. **The first strategy** involves extracting the fused Ara-Mental embeddings from learned using DJAE and

employing various models for classification (SVM, logistic regression and MLP).

2. **The second strategy** focuses on end-to-end training starting from the input texts (Arabic and Mental) passing by the joint Ara-Mental representation learning.

Results presented in Table 9 illustrate the different strategies applied to obtain the optimal Ara-Mental BERT model for Reddit dataset (D1) [20] classification.

Table 9 shows that the simple concatenation method achieved lower accuracy than the joint representation learning methods (joint autoencoder and DJAE). The simple concatenation gets 77.89% accuracy while the joint methods are in the 79–81% range. Between the joint representation methods, the DJAE model outperforms the basic joint autoencoder, achieving 81.12% vs 77.96% accuracy. This illustrates the benefits of the optimizations in the DJAE model. For end-to-end training, DJAE+MLP achieve the

Table 9 Test accuracy for mental health classification methods

Combination technique	End to end	Ara-Mental Representation + Classifiers		
		Fully connected MLP	SVM	Logistic regression
Simple Concatenation	77.89	77.89	79.23	78.83
Joint Autoencoder	77.96	79.17	79.03	78.29
DJAE	81.12	79.17	79.44	79.10

best accuracy of 81.12%. Furthermore, the DJAE representations with separate classifiers achieve slightly lower accuracy of 79.17–79.44% depending on the classifier used (MLP, SVM, logistic regression). Finally, the end-to-end deep joint autoencoder coupled with an MLP classifier achieved the best test accuracy of 81.12%. This demonstrates the benefits of joint representation learning optimized for the task, improving generalization.

Analyzing performance across different input types in Table 10 shows the combination of post text and titles results in the highest recall, precision, F1-score, and specificity. This highlights the importance of combining both content types for effective classification using the Ara-MentalBERT model.

Table 10 provides details about the precision, recall, F1 scores, and specificity per class using Ara-MentalBERT on posts, titles, and combined posts+titles. Performance is higher overall for posts+titles compared to posts or titles alone across most classes. Certain classes like bipolar disorder achieve very high precision (0.92–0.96), recall (0.96–0.99), F1 scores (0.95), and specificity (0.97–0.99) using posts+titles. This indicates Ara-MentalBERT is extremely effective at detecting bipolar disorder when provided sufficient text, with a very low false positive rate. Anxiety also exhibits strong performance with posts + titles, with precision of 0.83, recall of 0.75, F1 of 0.79, and specificity of 0.96. The model accurately identifies anxiety-related posts while maintaining a high ability to correctly classify non-anxiety cases. ADHD and None classes show good performance with posts + titles, suggesting the joint representations capably capture signals related to these conditions. The high specificity (0.96 for ADHD and 0.95 for None) indicates the model's strength in correctly identifying individuals without these conditions. Depression detection demonstrates more challenging, with lower precision (0.63–0.75), recall (0.52–0.72), and F1 (0.53–0.73) across all inputs compared to other classes. However, the specificity remains high (0.91–0.94), suggesting that while the model may miss some depression cases, it rarely misclassifies non-depression cases as depression. PTSD also has lower metrics for precision, recall, and F1-score, indicating difficulties

in identifying related posts. However, its high specificity (0.91–0.95) across all input types suggests the model is good at avoiding false positives for PTSD. Notably, specificity is consistently high across all classes and input types (ranging from 0.91 to 0.99). This is particularly important in a mental health context, as it indicates the model's strong ability to avoid misdiagnosing individuals who do not have a particular condition. The highest specificity values are observed for the posts + titles combination, further emphasizing the benefit of using both content types. In summary, the combination of posts and titles not only improves the model's ability to correctly identify mental health conditions (as shown by increased precision, recall, and F1 scores) but also enhances its capability to correctly identify individuals without these conditions (as evidenced by the high specificity). This balanced performance is crucial in mental health applications, where both accurate diagnosis and avoiding misdiagnosis are equally important.

The confusion matrices presented in Fig. 11 provides a comprehensive view of Ara-MentalBERT's performance on the Reddit Dataset (D1) across different input types: titles, posts, and the combination of posts and titles. These visualizations complement and reinforce the class-wise metrics detailed in Table 10, providing deeper insights into the model's classification capabilities.

A clear progression in accuracy is evident as we move from using only titles to posts, and finally to the combination of posts and titles. This progression is reflected in the increasing values along the diagonal of each matrix, indicating improved correct classifications across all mental health categories. Bipolar disorder consistently demonstrates exceptional performance across all input types, with the combination of posts and titles yielding the best results (245 correct classifications out of 248 instances). This aligns with the outstanding F1-score of 0.95 noted in Table 10, suggesting that Ara-MentalBERT excels at identifying unique linguistic patterns associated with bipolar disorder discussions.

Depression shows the most significant improvement when transitioning from titles alone to the combination of posts and titles, with correct classifications increasing from 128 to 183. Notably, the misclassifications with related conditions

Table 10 Results: Ara-MentalBERT Class-wise precision, recall, F1-score, and specificity across different input types for Reddit dataset (D1) [20]

Class	Posts				Titles				Posts + titles			
	P	R	F1	S	P	R	F1	S	P	R	F1	S
adhd	0.70	0.67	0.69	0.94	0.68	0.54	0.60	0.95	0.80	0.75	0.77	0.96
anxiety	0.81	0.72	0.76	0.96	0.57	0.60	0.58	0.91	0.83	0.75	0.79	0.96
bipolar	0.94	0.96	0.95	0.98	0.86	0.90	0.88	0.97	0.92	0.99	0.95	0.99
depression	0.63	0.68	0.66	0.92	0.54	0.52	0.53	0.91	0.75	0.72	0.73	0.94
ptsd	0.63	0.71	0.67	0.91	0.64	0.72	0.68	0.91	0.72	0.84	0.77	0.95
none	0.79	0.73	0.76	0.96	0.65	0.65	0.65	0.93	0.83	0.79	0.81	0.95
Average	0.75	0.75	0.75	0.94	0.65	0.66	0.65	0.93	0.81	0.81	0.81	0.96

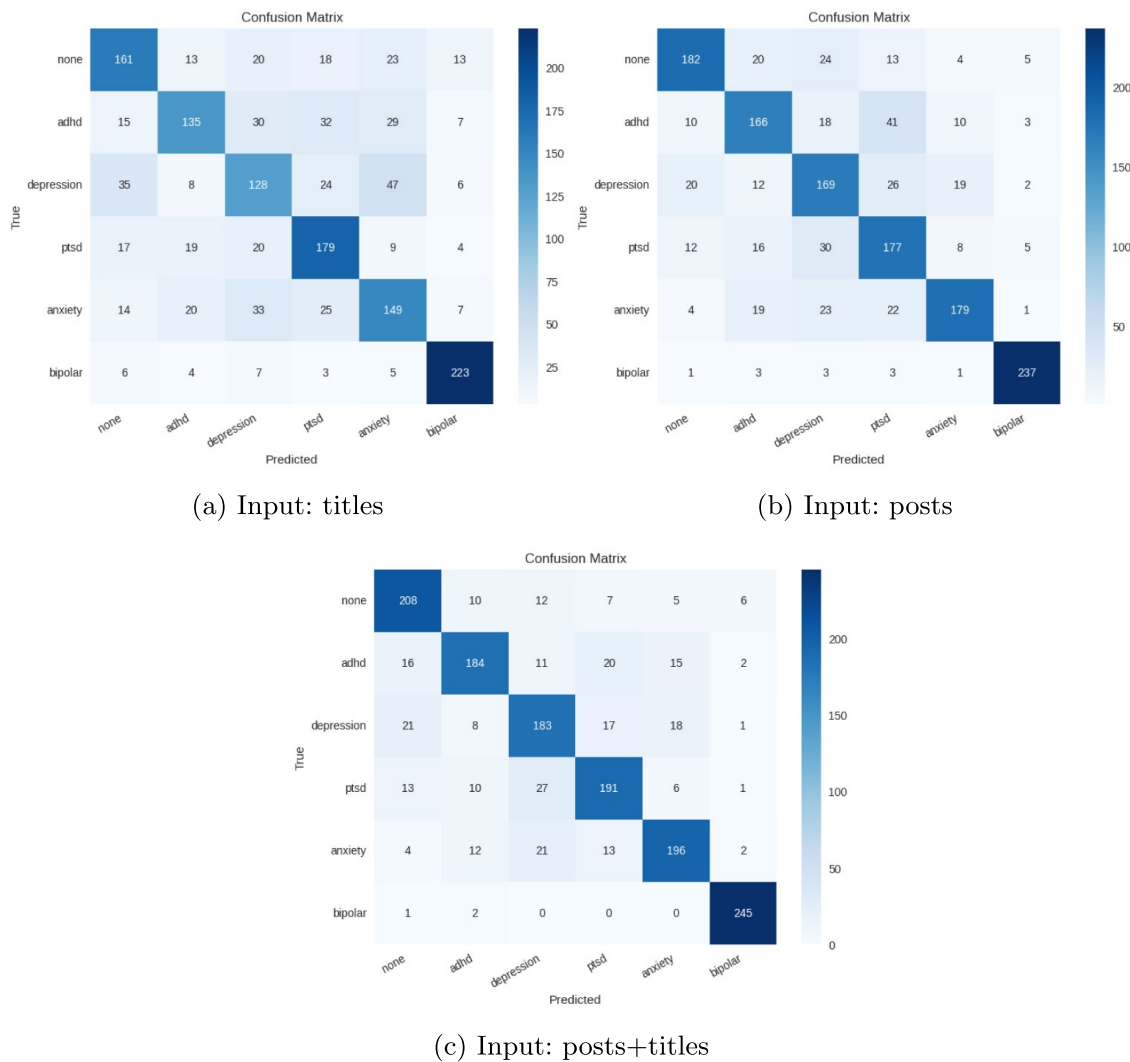


Fig. 11 Ara-MentalBERT: Confusion Matrices of Reddit Dataset (D1)

such as PTSD and anxiety decrease substantially, addressing a key challenge in distinguishing between these often comorbid conditions in textual data.

ADHD classification exhibits considerable enhancement, with correct identifications rising from 135 using only titles to 184 with posts and titles combined. The confusion with PTSD and anxiety reduces noticeably, indicating the model's improved ability to discern ADHD-specific language patterns when provided with more comprehensive input. Anxiety and PTSD also show marked improvement with the inclusion of both posts and titles. Anxiety correct classifications increase from 149 to 196, while PTSD improves from 179 to 191. In both cases, there is a notable reduction in misclassifications with related conditions, particularly depression.

The “none” category, representing posts without specific mental health content, demonstrates steady improvement

across input types, with 208 correct classifications when using posts and titles combined, compared to 161 with titles alone. This indicates enhanced ability to distinguish general discussions from mental health-related content when more context is provided.

Overall, these confusion matrices visually confirm the quantitative improvements observed in Table 10, underscoring the superior performance of the Ara-MentalBERT model when utilizing both posts and titles as input. The matrices clearly illustrate the reduction in misclassifications between related mental health conditions as more comprehensive textual information is provided, highlighting the model's increased capacity to capture nuanced differences in language use across various mental health discussions on social media platforms.

To evaluate the performance of the proposed Ara-MentalBERT + DJAE model, comprehensive experimentation

was conducted on the Arabic Twitter dataset (D2) [31] and the Arabic Reddit dataset (D1) [20]. The results obtained demonstrate the generalizability and robustness of our model across various mental health classification tasks.

Table 11 presents the class-wise precision, recall, F1-score, and specificity for the Arabic tweet classification task. Our Ara-MentalBERT model outperforms the baseline AraBERT and MentalBERT models in several categories, including FW (Feelings of worthlessness), SD (Sleep disorder), and SU (Suicidality). However, AraBERT exhibits higher macro-averaged and weighted-averaged metrics, indicating its strength as a general-purpose Arabic language model. The Ara-MentalBERT model strikes a balance between domain-specific and general language knowledge, achieving competitive results across most categories. The specificity scores presented in Table 11 provide additional insight into the model's performance. Notably, all categories demonstrate high specificity values, ranging from 0.96 to 0.99, with a macro average of 0.99. This indicates that the Ara-MentalBERT model excels at correctly identifying negative cases across all mental health categories. The consistently high specificity scores suggest that the model has a low false positive rate, which is crucial in mental health applications where misclassification could lead to unnecessary interventions or stigmatization. Categories such as FW (Feelings of worthlessness), LE (Loss of energy), SU (Suicidality), and WD (Weight disorder) show particularly high specificity (0.99), demonstrating the model's ability to accurately distinguish these critical mental health indicators from other categories. Even categories with relatively lower F1-scores, such as LM (Low mood) and PAR (Psychomotor agitation or retardation), maintain high specificity (0.96 and 0.97 respectively), indicating that while the model may sometimes miss positive cases in these categories, it rarely misclassifies negative cases. This balance of performance metrics underscores the Ara-MentalBERT model's potential

as a reliable tool for mental health screening and analysis in Arabic social media contexts.

The confusion matrix presented in Fig. 12 provides a detailed view of the Ara-MentalBERT model's performance on the Arabic Tweets Depression Dataset (D2), complementing the class-wise metrics in Table 11. The matrix reveals strong diagonal elements across most categories, indicating good overall classification accuracy for various depression symptoms. Notably, the model excels in identifying "Diminished ability to think or concentrate" (50 out of 55 instances correctly classified) and "Weight disorder" (37 out of 37 instances correctly identified), demonstrating its effectiveness in recognizing distinct linguistic patterns associated with these symptoms.

The model shows robust performance in detecting "Suicidality" (26 out of 28 instances correctly classified) and "Feelings of worthlessness" (22 out of 23 instances correctly identified). This is particularly crucial given the severity and clinical importance of these symptoms in depression diagnosis and intervention.

"Sleep disorder" and "Loss of energy" categories also show strong performance, with 18 out of 19 and 15 out of 18 instances correctly classified, respectively. This indicates the model's capability to capture subtle linguistic cues related to these physical symptoms of depression.

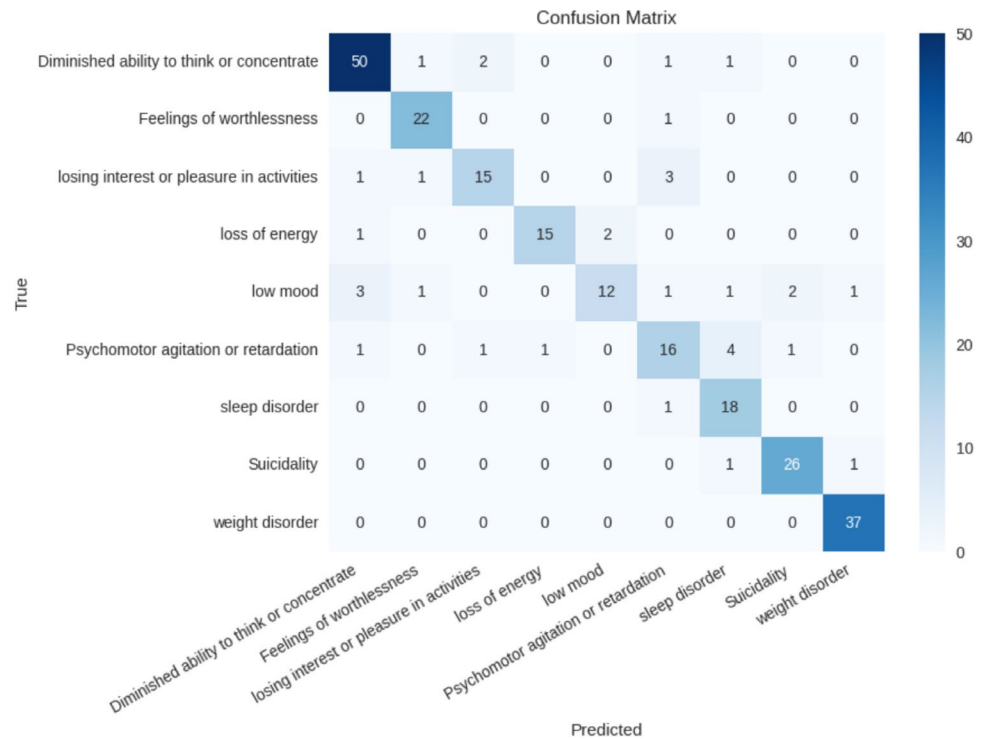
However, the confusion matrix also highlights some areas of challenge. "Low mood," a core symptom of depression, shows more misclassifications compared to other categories, with only 12 out of 21 instances correctly identified. This suggests that expressions of low mood may share linguistic features with other depressive symptoms, making it more challenging for the model to distinguish. There is some confusion between "Psychomotor agitation or retardation" and "Sleep disorder" (4 instances), as well as between "Losing interest or pleasure in activities" and "Psychomotor agitation or retardation" (3 instances). This indicates the complexity of differentiating between these related symptoms based on textual data alone.

Overall, the confusion matrix demonstrates the Ara-MentalBERT model's strong capability in identifying and distinguishing between various depressive symptoms in Arabic tweets. It performs exceptionally well on certain critical categories like suicidality and cognitive symptoms, while also highlighting areas for potential improvement, particularly in distinguishing between closely related emotional and behavioral symptoms of depression. This visualization provides valuable insights into the model's strengths and limitations, offering directions for future refinements to enhance its performance in mental health classification tasks on Arabic social media data.

In summary, Ara-MentalBERT utilizes joint representations very effectively for certain disorders like bipolar and anxiety. The results show relative strengths and weaknesses

Table 11 Results: Ara-MentalBERT Class-wise precision, recall, F1-score, and specificity for the Arabic Tweets Depression Dataset (D2) [31] classification

Category	P	R	F1	S
DATC	0.89	0.91	0.90	0.97
FW	0.88	0.96	0.92	0.99
LE	0.94	0.83	0.88	0.99
LIPA	0.83	0.75	0.79	0.98
LM	0.86	0.57	0.69	0.96
PAR	0.70	0.67	0.68	0.97
SD	0.72	0.95	0.82	0.97
SU	0.90	0.93	0.91	0.99
WD	0.95	1.00	0.97	0.99
Macro avg.	0.85	0.84	0.84	0.99

Fig. 12 Ara-MentalBERT: Confusion Matrice of the Arabic Tweets Depression Dataset (D2)**Table 12** Results of different models for mental health disorders classification (average results obtained with Reddit dataset (D1) [20])

Model	Titles			Posts			Titles + posts		
	P	R	F1	P	R	F1	P	R	F1
BERT	62.61	61.90	61.75	64.05	63.04	62.67	70.30	69.22	69.22
ClinicalBERT	56.13	55.11	55.27	60.33	59.88	59.84	68.20	66.94	67.12
BioBERT	58.94	57.80	57.98	62.78	62.70	62.38	72.43	72.11	72.06
RoBERTa (English corpus)	65.28	63.78	63.84	73.42	72.85	72.86	77.27	76.75	76.88
MentalBERT	69.09	68.01	67.83	76.28	75.74	75.83	81.74	81.18	81.19
AraBERT	61.22	60.42	60.58	71.61	70.03	70.26	74.91	74.13	74.23
RoBERTa (Arabic corpus)	20.90	25.20	21.20	33.77	30.78	27.12	36.07	37.90	35.08
Ara-MentalBERT + DJAE	65.48	65.52	65.32	75.10	74.60	74.71	81.09	81.12	81.10

for certain classes and inputs to guide future targeted optimization of Ara-MentalBERT. The model performs well on this text classification task, especially when combining both posts and titles. However, there are still opportunities for improvement, particularly in challenging classes like depression.

Discussion

Tables 12 and 13 present a comprehensive comparative analysis with state-of-the-art models for both the Reddit dataset (D1) and Twitter dataset (D2) classification tasks. Across all input configurations, our Ara-MentalBERT + DJAE model consistently outperforms the baseline models, including: BERT, RoBERTa, ClinicalBERT, BioBERT, and AraBERT model.

Table 13 Results of different models for mental health disorders classification (average results obtained with Tweet dataset (D2) [31])

Model	P	R	F1
AraBERT	94.41	93.47	93.56
MentalBERT	86.57	86.53	86.27
Ara-MentalBERT + DJAE	86.31	86.12	85.78

- The results demonstrate that specialized models (e.g., MentalBERT for mental health-related tasks) tend to outperform general-purpose models (e.g., BERT) on Reddit dataset.
- For the Reddit dataset, our model achieves the highest accuracy of 81.12%, recall of 81.12%, precision of

81.09%, and F1-score of 81.10% on the titles+posts input. This superior performance highlights the advantages of multi-domain fusion and the ability to leverage longer text contexts. While MentalBERT performs better on English posts, combining it with AraBERT using the DJAE framework improves the accuracy of AraBERT from 70.03% to 74.60%. For the Tweet dataset (D2), our proposed model, Ara-MentalBERT + DJAE, achieves an accuracy of 86.12%, a recall of 86.12%, a precision of 86.31%, and an F1-score of 85.78% on Arabic tweets. This demonstrates that our model maintains competitive performance across different datasets and languages, particularly highlighting its robustness in handling diverse text formats and its effectiveness in multi-lingual and multi-domain contexts.

- The superior results achieved by our Ara-MentalBERT + DJAE model validate the promise of transfer learning in the mental health domain. By leveraging multi-domain knowledge and effective representation learning, our model demonstrates its capability to accurately classify mental health-related content across various input sources and languages. These findings pave the way for further advancements in developing robust and reliable mental health monitoring systems, ultimately contributing to better support and intervention strategies for individuals in need.

Conclusion

In this research, we have successfully developed an effective automatic detection method for mental health disorders from user-generated content by strategically integrating the knowledge from AraBERT and MentalBERT through transfer learning techniques and a model joint approach. The proposed Ara-MentalBERT + DJAE model demonstrates superior performance, outperforming other state-of-the-art models in mental health classification tasks across various input configurations and languages. Our model achieved impressive results on both the Reddit Dataset (D1) and the Arabic Tweet Dataset (D2). For the Reddit Dataset, using combined titles and posts as input, our model attained an accuracy of 81.12%, precision of 81.09%, recall of 81.12%, and an F1-score of 81.10%. These results showcase the model's ability to effectively leverage longer text contexts. On the Arabic Tweet Dataset, our model demonstrated robust performance with an accuracy of 86.12%, precision of 86.31%, recall of 86.12%, and an F1-score of 85.78%, highlighting its effectiveness in handling Arabic language data, particularly for tweet classification tasks.

By employing the strengths of both Arabic and mental health domains, our model achieves a remarkable results for mental disorder classification.

Despite the challenges posed by limited Arabic language resources, our approach of dataset translation serves as an important initial step, enabling the development of a robust model capable of handling both English and Arabic mental health-related content. The superior performance of our proposed model validates the promise of transfer learning in the mental health domain and highlights the effectiveness of multi-domain fusion and joint representation learning. Although our research has yielded promising results, there is still room for further improvement and development. Future work can explore the potential of our model by investigating the integration of our proposed model into real platforms for psychological assistance could aid psychologists in their decision-making processes, ultimately enhancing the quality of mental health support and intervention strategies. Furthermore, the exploration of other Transformer-based models, such as pre-trained GPT models, may unveil new avenues for improving the automatic detection of mental health disorders from user-generated content.

Appendix: Hyperparameters Adjustment

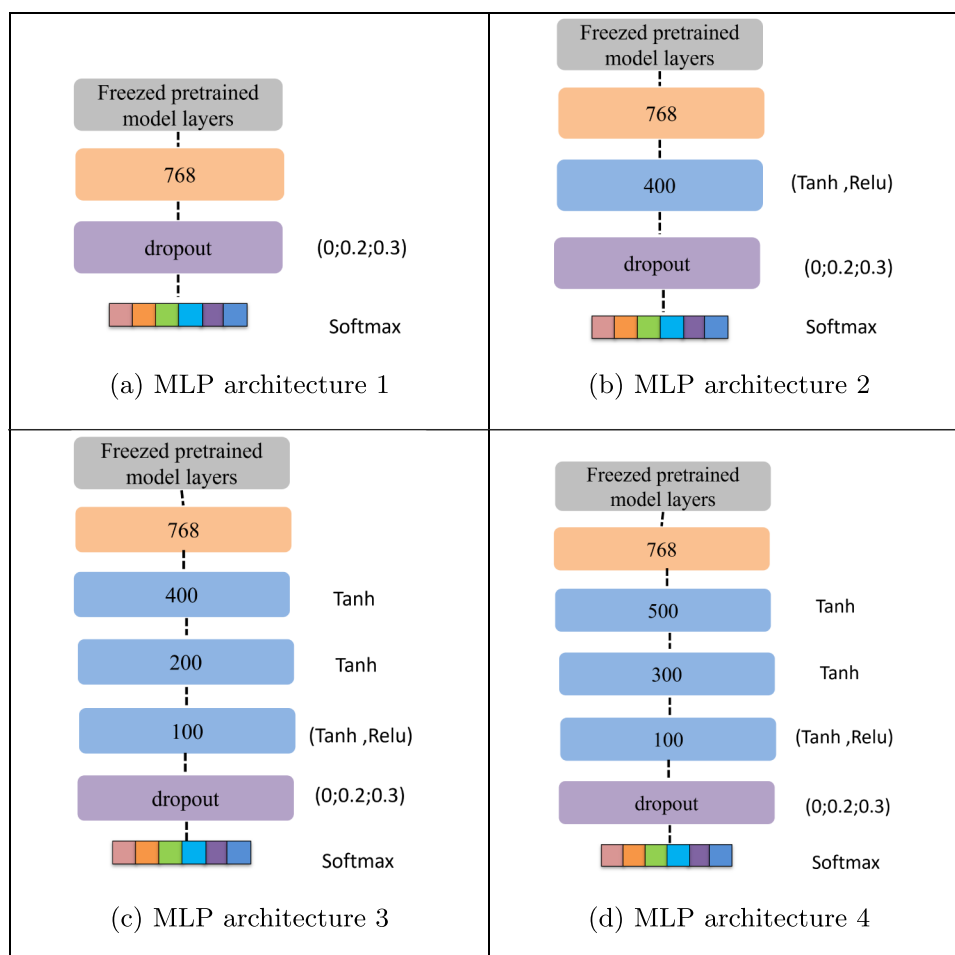
In this appendix, we describe the process of hyperparameter adjustment for the pre-trained language models AraBERT, MentalBERT, and the proposed Ara-MentalBERT model. The goal is to optimize the performance of these models on the task of mental health disorder classification from text. We conduct extensive experiments to find the optimal hyperparameter configurations, using the Reddit mental health dataset for validation.

AraBERT Hyperparameters Adjustment

The pre-trained AraBERTv02-base-uncased is used as our pre-trained model from hugging face library. In this section, we present the different configurations used for fine-tuning the AraBERT model, including the selection of hyperparameters, followed by ensuring results obtained on the test datasets. During the fine-tuning process, a classification token [CLS] is inserted into the input sequences and the corresponding output vector is used as the feature input to the classifier. Fully-connected layers with softmax output are added for prediction, with only these classifier layers fine-tuned.

To find the best-fitted architecture for fine-tuning AraBERT, we conduct hyperparameter adjustment using the validation dataset. The key hyperparameters we adjusted are (i) MLP implementation, (ii) epochs number and (iii) batch size.

1. **MLP implementation:** We add fully connected layers on top of the pre-trained AraBERT model to create a

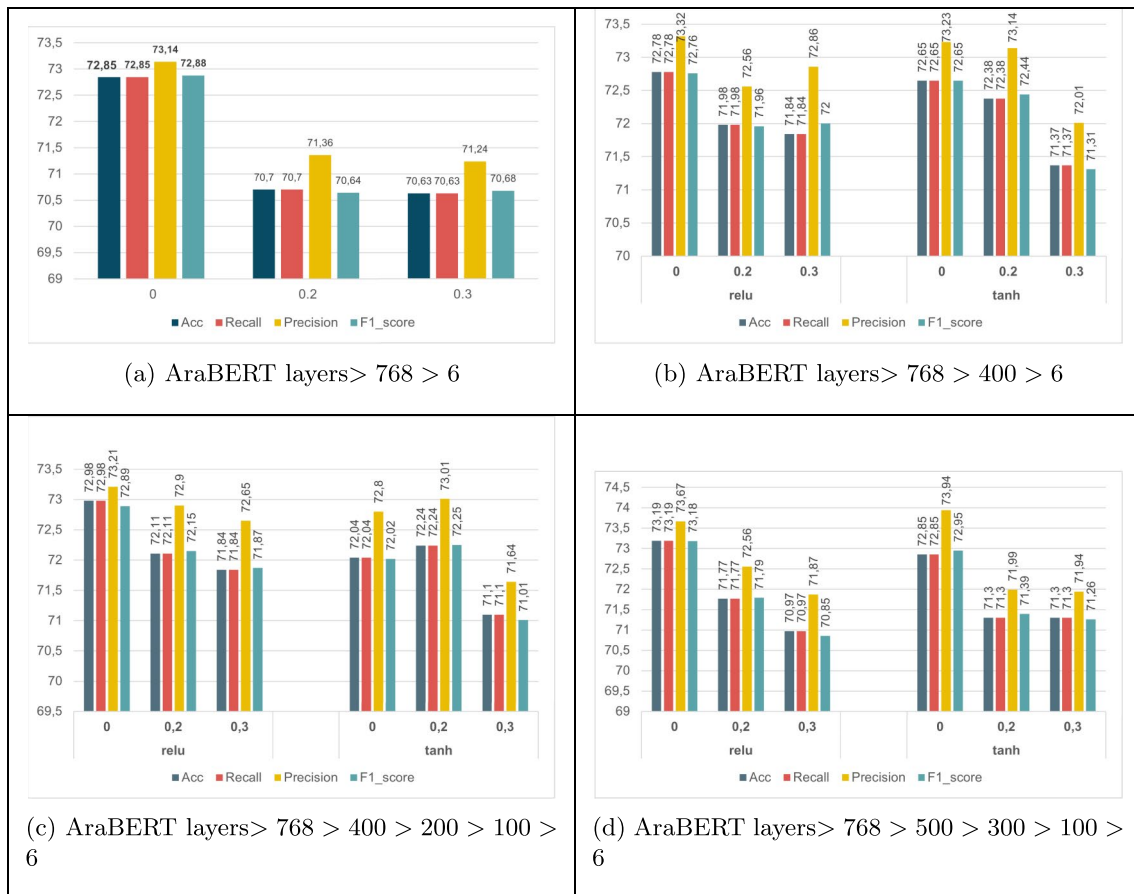
Table 14 MLP implementation for pre-trained model fine-tuning

classifier tailored for our downstream task. We experiment with different numbers and sizes of fully connected layers. While the pre-trained layers are frozen, the classifier layers consisting of MLPs are trained. In all experiments, the number of epochs are fixed at 10 and the learning rate is set to $1e-6$. The goal is to determine the optimal number of layers, neurons at each layer, and dropout rate for a dropout layer preceding the the decision layer. Four configurations are evaluated (see Table 14). The goal of this experiment is to evaluate different MLP architectures and hyperparameters to find the best performing model for fine-tuning AraBERT on a downstream task.

- **Architecture 1** depicted in Table 14a, utilize AraBERT with all layers frozen extracting 768 features which are then passed through a dropout layer with rates of 0, 0.2, and 0.3. This aims to assess the effect of dropout on frozen AraBERT features. The results on the validation set are shown in Table 15a. Applying a dropout layer decreases performance compared to directly classifying from the pooler features. From

the graph (Table 15a), it appears that increasing the dropout rate from “0” to “0.2” and “0.3” leads to a decrease in all four metrics (Accuracy, Recall, Precision, F1-Score). This suggests that a higher dropout rate negatively impacts the performance of the model. Based on validation accuracy, Architecture 1 without dropout is preferred. Further analysis on the impact of dropout and regularization for AraBERT fine-tuning will be examined in future experiments.

- In **Architecture 2**, dense layer with 400 units are added after AraBERT before the dropout and softmax layer (see Table 14b). This analyses the impact of activation functions (ReLU and Tanh) on dropout regularization (0, 0.2, 0.3). The validation results are presented in Table 15b. Experiments demonstrate that ReLU activation slightly outperforms Tanh, and increasing dropout generally decreases performance. The best validation accuracy is achieved using ReLU activation and 0.2 dropout with Accuracy. This configuration from Architecture 2 will be taken forward for further analysis and will be compared against other proposed architectures.

Table 15 MLP results for AraBERT model fine-tuning

- In **Architecture 3**, the full AraBERT model is frozen and additional dense classifier layers of different sizes (400 > 200 > 100) are added. Both ReLU and Tanh activations are tested for the 100 unit layer before the dropout and softmax classification illustrated in Table 14b. This experimental design aims to evaluate the impact of ReLU and Tanh activation for layer of 100 units shape and the effect of dropout rate (0, 0.2, 0.3). As shown in Table 15c, using ReLU activation for the 100 unit layer before dropout achieve better performance than Tanh activation. Specifically, ReLU with no dropout attains the highest validation accuracy of 72.98%, recall of 72.98%, precision of 73.21%, and F1-score of 72.89%. Applying a small dropout of 0.2 still produces competitive results with ReLU activation, indicating that some regularizations are possible without significantly degrading metrics. However, increasing the dropout to 0.3 leads to a notable reduction in all evaluation metrics for both ReLU and Tanh activation functions. Overall, ReLU activation performs better with the fully connected layers added to AraBERT. While ReLU without dropout achieves the best vali-

ation results, a moderate dropout of 0.2 provides an acceptable accuracy (72.98%). These results align with those from Architecture 2, further demonstrate the superiority of ReLU over Tanh for optimized classification layers in these model architectures.

- In **Architecture 4**, AraBERT is frozen and classifier layers of 500, 300, and 100 units are added. Both tanh and ReLU activation are tested before the dropout and softmax output. This architecture evaluates larger 500 unit layer compared to Architecture 3. As shown in Table 15d, the configuration with ReLU and no dropout achieve the best validation performance with accuracy of 73.19%, recall of 73.19%, precision of 73.67%, and F1-score of 73.18%.

This demonstrates that the combination of fully connected classifier layers and ReLU activation improve upon the results from Architecture 3. However, applying a dropout of 0.2 or 0.3 result in a significant decline in evaluation metrics. In summary, the results of Architecture 4 are consistent with previous results. In fact using MLP architecture (500 > 300 > 100 > 6) associated with ReLU activation and without dropout. We obtain

Table 16 AraBERT results obtained with different epochs numbers

Epochs	Accuracy (%)	Recall (%)	Precision (%)	F1 (%)
15	72.58	72.58	73.31	72.63
20	71.77	71.77	72.75	71.83
50	72.38	72.38	73.01	72.46

The bold values represent the best obtained results for each respective evaluation metric

the best results compared to other assessed MLP architectures.

2. **Epochs number:** The number of passes through the full training dataset. We test training for different numbers of epochs to find the ideal amount for convergence without overfitting. Following the MLP architecture experiments, Architecture 4 ($500 > 300 > 100 > 6$) with ReLU activation and without dropout is selected as the optimal model configuration. Further hyperparameter tuning on number of training epochs is then conducted to determine the best epoch setting. The AraBERT classifier is trained for 15, 20, and 50 epochs while keeping other hyperparameters constant. The obtained results using validation dataset are in Table 16. Based on the validation metrics, training for 15 epochs achieves the highest accuracy and F1 score. Using more epochs leads to overfitting and worse performance.
3. **Batch size:** The number of samples propagated through the network before the parameters are updated. We try different batch sizes to optimize computational efficiency and generalization. After determining the optimal number of training epochs, experiments are conducted to tune the batch size hyperparameter. The AraBERT classifier is trained using batch sizes of 16, 32, 50, and 100 with 15 epochs. The obtained results using validation dataset are detailed in Table 17. Increasing batch size from 16 to 32 improves both accuracy and F1-score. However, further increase to 50 and 100 decreases performance. According to achieved results, the optimal batch size is 32.

In conclusion, the fully tuned AraBERT model employs an additional MLP with architecture contains ($500 > 300 > 100 > 6$) trained for 15 epochs with a batch size of 32. This optimal configuration will be utilized for the final evaluation on the test dataset.

MentalBERT Hyperparameters Adjustment

Similarly, we adjust the hyperparameters, just as we do when fine-tuning AraBERT. Many experiments are conducted

Table 17 AraBERT results obtained with varying batch size

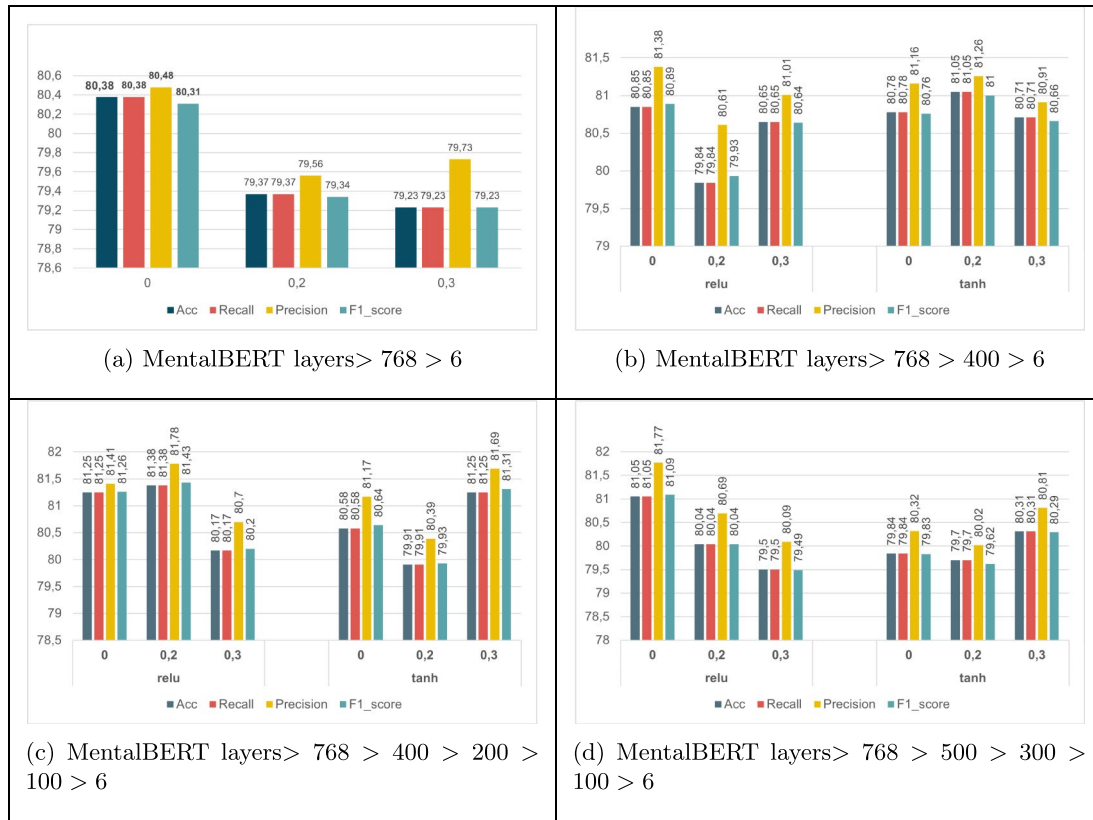
Batch size	Accuracy (%)	Recall (%)	Precision (%)	F1 (%)
16	72.58	72.58	73.31	72.63
32	73.25	73.25	74.04	73.35
50	72.58	72.58	72.92	72.58
100	72.18	72.18	72.76	72.21

The bold values represent the best obtained results for each respective evaluation metric

to optimize the fine-tuning of MentalBERT on the Reddit dataset to generate contextualized representations for mental health classification. The pretrained MentalBERT parameters are kept frozen and classifier layers are added for prediction. The [CLS] token embedding from MentalBERT provide the input representation for the classification task.

1. **MLP implementation:** The hyperparameter configurations for MentalBERT fine-tuning are presented in Table 14, showcasing four different configurations which are evaluated on the validation dataset.

- In **Architecture 1**, Frozen MentalBERT utilizes with one dropout layer (rated as 0, 0.2 and 0.3) and the decision layer. The obtained results depicted in Table 18a reveal that the configuration without dropout layer achieve the best performance, with 80.38% accuracy, 80.38% recall, 80.48% precision, and 80.31% F1-score. The introduction of dropout leads to a decrease in performance, indicating that regularizing the pooler representation (768 dimension embedding) negatively affects transfer learning from the pretrained model. Therefore, the optimal design for Architecture 1 is to direct classification the MentalBERT output embeddings without dropout.
- **Architecture 2**, involves freezing the pretrained MentalBERT layers and adding a layer with 400 units, one dropout layer and decision layer with softmax activation. Both ReLU and Tanh activations are tested, along with varying dropout rates for regularization. As shown in Table 18b, the configuration with Tanh activation and 0.2 dropout achieve the best validation performance, with 81.05% accuracy, 81.05% recall, 81.26% precision and 81.00% F1-score. For a dropout of 0.2 to 0.3, the tanh activation outperforms ReLU in terms of all evaluation metrics. This indicates that the Tanh function is a better fit for the classification layer applied to MentalBERT embeddings. Moderate 0.2 dropout provides a slight benefit with Tanh activation. While, higher 0.3 dropout decreases performance due to

Table 18 MLP results for MentalBERT model fine-tuning

excessive regularization. In summary, Architecture 2 demonstrates that adding classification depth with a layer of (400 units and tanh activation) improves upon directly classifying from the MentalBERT pooler in Architecture 1 with lightweight regularization further enhancing the results.

- Architecture 3** aims to evaluate the effect of increased classifier depth by appending additional dense layers of 400, 200, and 100 units after the frozen MentalBERT layers. Both ReLU and tanh activations are tested for the 100 unit layer before the output. As shown in Table 18c, the configuration with ReLU activation and 0.2 dropout achieve the best validation performance, with 81.38% accuracy, 81.38% recall, 81.78% precision and 81.43% F1-score. The ReLU activation outperforms Tanh across all dropout levels in the deeper 100 unit layer. This demonstrates that ReLU is better suited than tanh in deeper classifier architectures appended to MentalBERT. Applying a small 0.2 dropout rate provided a small benefit in regularization and improves the results of the ReLU network. However, a high dropout rate of 0.3 decline the performance due to excessive regularization. Overall, the Architecture 3 experiments show that increasing classification depth

with additional ReLU activated layers enhances performance compared to Architecture 2.

- Architecture 4**, explores the impact of even larger classifier layers by evaluating dense layers of 500, 300, and 100 units added to the frozen MentalBERT layers with both ReLU and Tanh activations. The main objectives are to evaluate a layer 500 units larger compared to architecture 3, to evaluate the effect of depth with layers 300 and 100 and to analyze the influence of dropout regularization.

As shown in Table 18c, the configuration with ReLU activation and no dropout achieves the best validation accuracy of 81.05% and F1-score of 81.09%. However, all metrics decrease with increased dropout, indicating over-regularization. Tanh activations performs less effectively than ReLU across all dropout rates in the large Architecture 4 highlighting the applicability of ReLU for the deeper MLP layers added to MentalBERT. In summary, Architecture 3 with frozen MentalBERT layers and added classifier layers of 400 units (tanh activation), 200 units (tanh activation), and 100 units (ReLU activation) achieves the best validation performance. Applying a dropout rate of 0.2 before the output layer results in the top accuracy and F1-score. Based on

Table 19 MentalBERT Architecture 3 validation results for different training epochs

Epochs	Accuracy	Recall	Precision	F1_score
15	81.05	81.05	81.54	81.07
20	80.65	80.65	81.12	80.65
50	80.24	80.24	80.77	80.26

The bold values represent the best obtained results for each respective evaluation metric

Table 20 MentalBERT results obtained with varying batch size

Batch size	Accuracy	Recall	Precision	F1 Score
16	80.65	80.65	80.97	80.63
32	80.71	80.71	81.32	80.78
50	81.65	81.65	81.64	81.61
100	80.11	80.11	80.50	80.15

The bold values represent the best obtained results for each respective evaluation metric

the architecture experiments and hyperparameter tuning results, this configuration provides the most effective method of employing the pretrained MentalBERT representations and classifier depth for mental text classification.

2. **Epochs number:** After determining Architecture 3 as the optimal model configuration based on validation dataset, further experiments are conducted to find the optimal number of training epochs. Keeping other hyperparameters fixed, Architecture 3 is trained for 15, 20, and 50 epochs. The results on validation dataset are shown in Table 19. From Table 19, we can notice that training for 15 epochs achieves the best accuracy and F1-score. Thus, increasing the epochs number leads to overfitting, resulting decreased performance. Hence, the optimal number of training epochs for the MentalBERT classifier is determined to be 15. This configuration will be used for further tuning batch size and final model evaluation.
3. **Batch size:** Further hyperparameter tuning experiments are conducted to determine the ideal batch size using the chosen Architecture 3 with 15 training epochs. The MentalBERT classifier is trained with batch sizes of 16, 32, 50, and 100 samples. The results on validation dataset are presented in Table 20. Based on the obtained results, a batch size of 50 achieves optimum performance. Increasing the batch size to 100 adversely impacts results, while a smaller batch size also leads to sub-optimal results. Consequently, a batch size of 50 samples strikes a balance between computational efficiency and effectiveness for fine-tuning MentalBERT.

Table 21 Deep Joint autoencoder validation results for different training epochs

Epochs	Accuracy	Loss
15	78.16%	1.1911
50	81.05%	1.1034
100	79.84%	1.1303

The bold values represent the best obtained results for each respective evaluation metric

Ara-MentalBERT Hyperparameters Adjustment

We train a deep joint autoencoder (DJAЕ) to learn a fused representation named AraMental embedding from the embeddings obtained by fine-tuning both AraBERT and MentalBERT models. The objective is to integrate information from both models into a joint embedding space, enhancing the model's ability to capture nuanced features related to mental health disorders from Arabic text. As AraBERT and MentalBERT are pre-trained on different domains, standard Arabic text and English mental health forums respectively, combining them through a deep joint autoencoder requires some parameter tuning to find the optimal fusion. In the following section, we will explain the various experiments conducted for adjusting the hyperparameters of Ara-MentalBERT.

To create an optimal text representation for mental health analysis, we tune **deep joint auto-encoder hyperparameters**. Through iterative experiments adjusting these hyperparameters, we aim to arrive at an Ara-Mental BERT model that maximizes prediction accuracy on mental health datasets. The number of training epochs is a key hyperparameter when optimizing the deep joint autoencoder architecture for fusing Ara-Mental sentence representations. As shown in Table, the DJAE is trained for 15, 50, and 100 epochs while keeping other factors constant. The validation accuracy and overall reconstruction loss are compared (Table 21).

It is observed that 50 epochs achieve the best accuracy of 81.05% and the lowest loss of 1.1034. Using fewer epochs results in underfitting, with higher loss and reduced accuracy. Conversely, training DJAE for 100 epochs causes overfitting, increasing loss and decreasing accuracy.

The training loss over epochs for the deep joint autoencoder fusion model provides information about Ara-Mental representation optimization. As depicted in the loss curves (see Fig. 13), the total loss, AraBERT-specific loss, and MentalBERT-specific loss all decrease as epochs increase from 15 to 50 to 100.

The lowering losses indicate improving the model fitting with fusion embeddings capturing linguistic information from both Arabic and Mental representations. However, losses of the autoencoder and AraBERT show larger decreases than

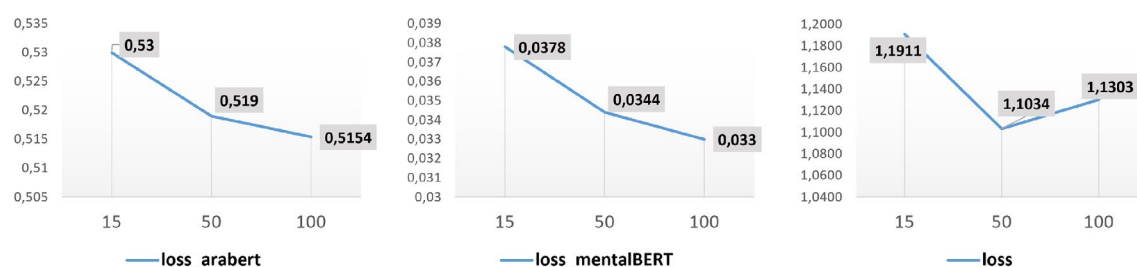


Fig. 13 Validation loss of deep joint auto-encoder over epochs

MentalBERT's, already starting from a lower baseline. This suggests that MentalBERT have well-optimized initial representations that continue to improve further with more epochs. In contrast, AraBERT's larger decrease shows a greater potential for enhancing its embeddings.

In summary, after 50 epochs the losses start increasing again, indicating overfitting of the autoencoder on language fusion. Therefore, 50 epochs achieve an ideal performance. At this point, multi-task learning shows promise in transferring encodings across Arabic and English to better recognize mental health disorders.

Author Contributions The contribution of all authors is equal.

Funding Not applicable.

Data Availability The datasets used in our study are available online via this link [amurark/mental-health-classification](https://github.com/amurark/mental-health-classification) GitHub repository for the Reddit dataset and the Mendeley data repository for the Twitter dataset, <https://data.mendeley.com/datasets/myrb2gky8w/1>.

Declarations

Conflict of Interest Not applicable.

Research Involving Human and/or Animals This article does not contain any studies with human or animal subjects performed by any of the authors.

Informed Consent Informed consent was obtained from all authors involved in the study.

References

- Kumar A, Sharma A, Arora A. Anxious depression prediction in real-time social data. In: International conference on advances in engineering science management & technology (ICAESMT)-2019. Uttaranchal University, Dehradun, India; 2019.
- Shatte AB, Hutchinson DM, Fuller-Tyszkiewicz M, Teague SJ. Social media markers to identify fathers at risk of postpartum depression: a machine learning approach. *Cyberpsychol Behav Soc Netw*. 2020;23(9):611–8.
- Chadha A, Kaushik B. Suicidal ideation from the perspective of social and opinion mining. In: *Proceedings of ICRIC 2019: recent innovations in computing*. Springer; 2020. p. 659–670.
- Chadha A, Kaushik B. Machine learning based dataset for finding suicidal ideation on twitter. In: *2021 Third international conference on intelligent communication technologies and virtual mobile networks (ICICV)*. IEEE; 2021. p. 823–828.
- Mbarek A, Jamoussi S, Hamadou AB. An across online social networks profile building approach: application to suicidal ideation detection. *Futur Gener Comput Syst*. 2022;133:171–83.
- Nova K. Machine learning approaches for automated mental disorder classification based on social media textual data. *Contemp Issues Behav Soc Sci*. 2023;7(1):70–83.
- LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436–44.
- Kim J, Lee J, Park E, Han J. A deep learning model for detecting mental illness from user content on social media. *Sci Rep*. 2020;10(1):11846.
- Cornn K. Identifying depression on social media. Stanford, CA: Department of Statistics Stanford University; 2020.
- Aragon ME, Lopez-Monroy AP, Gonzalez-Gurrola L-CG, Montes M. Detecting mental disorders in social media through emotional patterns-the case of anorexia and depression. *IEEE Trans Affect Comput*. 2023;14(1):211–22.
- Bouarara HA. Recurrent neural network (rnn) to analyse mental behaviour in social media. *Int J Softw Sci Comput Intell (IJSSCI)*. 2021;13(3):1–11.
- Shrestha A, Serra E, Spezzano F. Multi-modal social and psycho-linguistic embedding via recurrent neural networks to identify depressed users in online forums. *Netw Model Anal Health Inform Bioinform*. 2020;9:1–11.
- Almars AM. Attention-based bi-lstm model for arabic depression classification. *Comput Mater Contin*. 2022;71(2):3091–106.
- Coppersmith G, Leary R, Crutchley P, Fine A. Natural language processing of social media as screening for suicide risk. *Biomed Inf Insights*. 2018;10:1178222618792860.
- Zeberga K, Attique M, Shah B, Ali F, Jembre YZ, Chung T-S, et al. A novel text mining approach for mental health prediction using bi-lstm and bert model. *Comput Intell Neurosci*. 2022;2022:7893775.
- Albalawi Y, Nikolov NS, Buckley J, et al. Pretrained transformer language models versus pretrained word embeddings for the detection of accurate health information on arabic social media: Comparative study. *JMIR Format Res*. 2022;6(6):34834.
- Hu Y, Ding J, Dou Z, Chang H, et al. Short-text classification detector: a bert-based mental approach. *Comput Intell Neurosci*. 2022;2022:8660828.

18. Farruque N, Zaïane OR, Goebel R, Sivapalan S (2022) Deep-blues@ It-edi-acl2022: depression level detection modelling through domain specific bert and short text depression classifiers. In: Proceedings of the second workshop on language technology for equality, diversity and inclusion, pp 167–171
19. Putra AE, Maharani W. Depression levels detection through twitter tweets using roberta method. *J Inf Syst Res (JOSH)*. 2022;3(4):453–9.
20. Murarka A, Radhakrishnan B, Ravichandran S (2021) Classification of mental illnesses on social media using roberta. In: Proceedings of the 12th international workshop on health text mining and information analysis, pp 59–68
21. Chadha A, Kaushik B. A hybrid deep learning model using grid search and cross-validation for effective classification and prediction of suicidal ideation from social network data. *New Gener Comput*. 2022;40(4):889–914.
22. Phukan OC, Sharma R, Buduru AB. Trauma lurking in the shadows: a reddit case study of mental health issues in online posts about childhood sexual abuse; 2023. arXiv preprint [arXiv:2306.10338](https://arxiv.org/abs/2306.10338)
23. Grigore D-N, Pintilie I. Transformer-based topic modeling to measure the severity of eating disorder symptoms. *Working Notes of CLEF*, 2023;18–21.
24. Thushari PD, Aggarwal N, Vajrobol V, Saxena GJ, Singh S, Pundir A. Identifying discernible indications of psychological well-being using ml: explainable ai in reddit social media interactions. *Soc Netw Anal Min*. 2023;13(1):141.
25. Patel FP. Enhancing health tweet classification: an evaluation of transformer-based models for comprehensive analysis. PhD thesis, The University of Texas at Arlington; 2023.
26. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding; 2018. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805).
27. Antoun W, Baly F, Arabert HH: Transformer-based model for Arabic Language understanding. In: Proceedings of the 4th workshop on open-source Arabic corpora and processing tools, with a shared task on offensive language detection. European Language Resource Association, Marseille, France; 2020. p. 9–15.
28. Zeroual I, Goldhahn D, Eckart T, Lakhouaja A. Osian: open source international arabic news corpus-preparation and integration into the clarin-infrastructure. In: Proceedings of the fourth arabic natural language processing workshop; 2019. pp. 175–182.
29. Ji S, Zhang T, Ansari L, Fu J, Tiwari P, Cambria E. Mentalbert: publicly available pretrained language models for mental health-care; 2021. arXiv preprint [arXiv:2110.15621](https://arxiv.org/abs/2110.15621).
30. Ameer H, Jamoussi S, Hamadou AB. A deep neural network model for predicting user behavior on facebook. In: 2019 International joint conference on neural networks (IJCNN); 2019. IEEE. pp. 1–8.
31. Maghraby A, Ali H. Modern standard arabic mood changing and depression dataset. *Data Brief*. 2022;41: 107999. <https://doi.org/10.1016/j.dib.2022.107999>.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.