ISSN 2007-9737

Detection of Tendency to Depression through Text Analysis

Lauro Reyes-Cocoletzi*, J. Alejandro Aldama-Ramos, Alan Elias-Zapata,
Jorge Betancourt-González, Jésus Rojas-Hernández

Instituto Politécnico Nacional,
Unidad Profesional Interdisciplinaría de Ingeniería,
Mexico

{lreyesc, jrojashe}@ipn.mx,{jaldamar2000, aeliasz2000, jbetancourtg2000}@alumno.ipn.mx

Abstract. A project is proposed with the objective of detecting tendencies toward depression through text analysis, using Natural Language Processing technologies and Large Language Models (LLM). The development included several phases, such as the selection and preprocessing of English transcripts from the low-resource Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ) dataset [18, 19], as well as the training of models based on Transformer architectures, specifically Bidirectional Encoder Representations from Transformers (BERT), Robustly Optimized BERT Approach (RoBERTa), and Decoding-enhanced BERT with Disentangled Attention (DeBERTa). The results highlight the performance of the BERT fine-tuning model, which achieved better metrics compared to the other architectures evaluated (RoBERTa and DeBERTa fine-tuning models), with an average F1 score of 0.76 and a consistently high Receiver Operating Characteristic – Area Under the Curve (ROC-AUC) value > 0.82. This demonstrates its ability to balance precision and sensitivity, as well as identify linguistic patterns associated with depressive symptoms.

Keywords. Linguistic patterns, BERT, tendencies, depression, fine-tuning.

1 Introduction

Depression is a serious mental disorder that affects millions of people worldwide, in 2021, only in Mexico, 3.6 million people suffered from depression [11], early identification of this disorder is crucial to providing appropriate support and treatment. Natural language processing (NLP) has been used to predict depression, among other mental health conditions, from text sources such as social media posts, text messages, and journal entries [16]. Linguistic features, such as the use of pronouns and emotions, have proven to be useful in detecting and predicting depression.

First, depressed individuals tend to use more singular first-person words and pronouns in their writing, which may indicate higher levels of rumination, a constant focus on negative thoughts about themselves, or self-referential perseverance, often in a negative context [6]. Second, depressed individuals use more negative emotions and feelings in their writing [17]. In this context, the use of complementary interventions such as therapeutic writing has gained attention for its potential to help manage depression symptoms [21]. Likewise, the implementation of NLP methods in the field of emotion recognition is a key tool that can aid in the early identification of certain mental illnesses.

However, the aforementioned mechanisms are limited by the need for a rigorous clinical study, which includes adequate preparation of participants diagnosed with depression, as well as controls to prevent bias both from the patients and in the interpretation of the collected information.

For this reason, this research work uses a dataset obtained through a rigorous methodology carried out by health experts to collect textual information from patients with and without a diagnosis of depression (DAIC-WOZ). Given the clinical nature of the study, the dataset is relatively small, and due to the sensitivity of the data, no data augmentation was performed.

This paper is organized as follows: Section 2 describes the related work, Section 3 presents the proposed methodology, and Section 4 discusses the experimental results of the classification. Finally, Section 5 presents the conclusions and future work.

2 Related Work

A common approach to text analysis for detecting emotions and mental states is the use of personal communication media, separate from social networks. In many cases, mental health professionals use personal journals as a source of information.

In Anderson's work [1], personal journal entries are analyzed to identify the polarity (whether an emotion is positive or negative) of different texts. The dataset is organized by year, and the first stage involves preprocessing the entries. This includes removing empty paragraphs, breaking the text into smaller parts (tokenization), converting all text to lowercase, removing contractions, standardized date formats, and performing lemmatization.

In Li's work [8], ten years of personal writings are analyzed to determine the sentiment within them. A score is used to weigh positive, negative, or neutral sentiments. To generate this score, special attention is paid to the number of words and the time intervals between entries.

However, in the work of Vandana et al. [10], different architectures of Convolutional Neural Networks (CNN) and long-short term memory (LSTM) are proposed. The study is hybrid, as it is based on text responses provided by patients diagnosed with depression, as well as interviews (audio recordings) from which spectrograms are extracted. The network outputs assign binary labels Depressed and Not Depressed. The dataset used by Vandana is DAIC-WOZ, which is frequently used in automatic depression detection systems.

The DAIC-WOZ dataset, which includes audio, video, and textual information from individuals with depressive symptoms, was used strictly for research purposes. Access to this dataset was granted upon signing the official DAIC Data Usage Agreement Form [20], which outlines conditions for responsible use, data protection, and confidentiality. No personally identifiable information was used or disclosed, and all analyses were conducted in accordance with the terms stipulated by the dataset providers.

This dataset is highlighted due to its relevance and the validity of the study from which the information was obtained, making it the primary data source for this investigation.

Finally, it is also relevant to mention the existence of more advanced models based on the implementation of LLMs, such as GPT, which have demonstrated impressive performance in a wide range of natural language understanding tasks and have been explored in various recent studies [2, 5, 12, 9]. These models benefit from extensive pretraining on massive and diverse corpora, enabling them to capture complex linguistic and semantic patterns. However, their use requires substantial computational resources, including high-performance hardware such as GPUs or TPUs, as well as large-scale datasets to fine-tune them effectively.

Given the hardware constraints of the present research, a direct comparison with these models is not feasible. Moreover, the methodological approach of this study prioritizes the implementation of a lightweight and computationally efficient model that can still deliver meaningful results by identifying linguistic patterns within a clinically validated dataset, albeit limited in size. This focus allows for greater interpretability and applicability in real-world clinical or low-resource settings where computational power and access to extensive data are often restricted.

3 Proposed Methodology

The methodology consists of the following stages: preparation and preprocessing of only the textual information from the DAIC-WOZ dataset, tokenization, encoding of the correlation between the questions posed by the speaker and the responses provided by the participants in the analyzed questionnaires, and the generation of embeddings based on the construction of this question–answer correlation. The resulting embeddings are then utilized as input for three LLMs, to perform transfer learning, with an

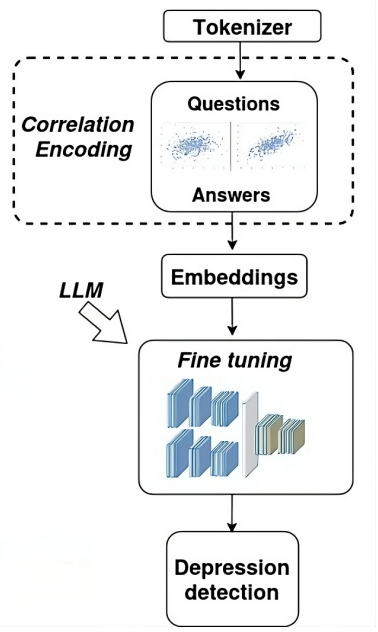

Fig. 1. General stages of the proposed methodology

emphasis on fine-tuning to identify behavioral patterns associated with depression.

Additionally, an experimental stage was included in which k-fold cross-validation was applied to ensure the robustness and generalizability of the models. This approach allowed for a more reliable estimation of the models' performance by mitigating the risk of overfitting and enabling the detection of consistent linguistic patterns across different data subsets.

Finally, the performance of depression detection is assessed using quantitative metrics to validate the classification accuracy of depressive tendencies. Figure 1 summarizes the stages of the methodology mentioned briefly, abstracting the most relevant blocks.

The relevant part of the methodology is clearly the LLMs: BERT, RoBERTa and DeBERTa. For example, in this work, BERT model and its pretrained variants are not, by themselves, ready for specific tasks such as binary classification of emotional or clinical states, as is the case in detecting depression from text. To adapt BERT to this type of supervised task, fine-tuning is performed by adjusting the pretrained model's parameters using the DAIC-WOZ dataset.

A key component of this adaptation involves modifying the classification head—the section of the model responsible for transforming BERT's contextual representations into task-specific predictions. In its original configuration, BERT does not include this layer for custom classification tasks. Therefore, it is necessary to replace or add a fully connected layer whose output size matches the number of classes in the target task.

For a binary classification task such as depression vs. non-depression, a linear layer is added on top of the contextual representation generated by BERT (typically of size 768 in the case of bert-base). This layer transforms the representation of the special token which summarizes the full input sequence—into a set of logits for each class. These logits are then passed through an activation function such as softmax (or sigmoid if using a single output unit) to compute the probability of belonging to each class.

This approach enables the reuse of the powerful architectures of BERT, RoBERTa, and DeBERTa, while efficiently adapting them to specific tasks such as the automatic identification of linguistic markers of depression in user-generated texts.

3.1 DAIC-WOZ Dataset Analysis

Training and evaluating models capable of identifying signs of depression in written texts require access to high-quality datasets. Among the various available corpora, DAIC-WOZ stands out as the most suitable for classifying texts into depression and non-depression categories.

DAIC-WOZ provides transcripts of semi-structured interviews conducted by a virtual conversational agent, specifically designed to detect signs of depression, anxiety, and post-traumatic stress disorder [3]. This dataset includes detailed clinical assessments based on the Patient Health Questionnaire-8 (PHQ-8) [7], ensuring the accuracy and relevance of the labels associated with the emotional states of participants [4].

Preprocessing is a crucial step in preparing data before training learning models. Various

Table 1. Score Structure in the DAIC-WOZ dataset [18]

Participant ID	PHQ-8 Binary	PHQ-8 Score
303	0	6 points
319	1	13 points
320	1	11 points
321	0	7 points
336	1	20 points

Table 2. Information example DAIC-WOZ Dataset [18]

Start time	Stop time	Speaker	Value
36.588	39.668	Ellie	Hi i am Ellie thanks for coming in today
39.88	43.788	Ellie	I was created to talk to people in a safe and secure environment
43.728	48.498	Ellie	think of me as a friend i do not judge i can not i am a computer
49.188	52.388	Ellie	i am here to learn about people and would love to learn about you
60.028	61.378	Ellie	how are you doing today

techniques were applied to clean, normalize, and structure the data to make it suitable for analysis. In this case, additional files containing the depression scores (PHQ-8) of each participant were also used, as shown in Table 1.

These scores are derived from clinical questionnaires and provide a quantitative measure of depression levels. According to specialists, a threshold was established: a score greater than or equal to 10 points indicates depression (label 1), while a score below 10 points indicates no depression (label 0).

A new column, PHQ-8 Binary, was created in the DataFrame to store these binary labels. Therefore, a merge was performed between the transcription DataFrame and the depression score DataFrame, using personId as the key. This facilitated the structured mapping of each question–answer pair to its respective depression label.

3.2 Processing of Transcriptions

Each transcription is stored in a separate file, where each row represents a line of dialogue containing information about the speaker ('Speaker') and the spoken text ('value'), as shown in Table 2. The transcripts feature two main interlocutors: Ellie, a virtual interviewer who poses the questions, and the participant, who serves as the interviewee and provides the answers.

Each transcription was processed line by line (construction of Question-Answer Pairs), when the Speaker is Ellie, the corresponding Value is considered a question and temporarily stored. The subsequent lines where the Speaker is the participant are concatenated to form the response to the previous question. This process continues until another intervention from Ellie is encountered, marking the beginning of a new question. Special cases, such as interruptions or empty lines, were handled to ensure the coherence of the question-answer pairs [18].

The extracted question and answer pairs were stored along with a unique participant identifier, personId. A DataFrame was constructed with the following columns: personId, question, and answer. Rows with empty questions or answers were removed. Additionally, a filter was applied to select only questions that begin with common interrogative words ("where," "when," "how," "why," "are," "what," "do," "have," "can," "did," "is," "could," "so," "tell," "who," "has").

This process structured the transcriptions into a suitable format for analysis, resulting in the question and answer columns, which represent the interactions between the interviewer and the participant. At this stage, from the 189 sessions in the original DAIC-WOZ dataset where 69.84% (132 sessions) corresponded to non-depression and 30.16% (57 sessions) corresponded to depression a total of 8,703 samples were obtained, with 68.10% (5,927 samples) classified as non-depression and 31.90% (2,776 samples) as depression.

Due to the percentage imbalance in the samples, random oversampling with replacement was applied to the minority class (depression) to equalize the number of samples with the

majority class (non-depression). Once the dataset was randomly balanced, it was split into specific proportions: 70% for training, 14% for validation, and 16% for testing, ensuring that each subset maintained a balanced representation of both classes. This process is detailed in the following sections.

3.3 Balanced Dataset

A random shuffle was performed on the balanced dataset to reorganize the samples before splitting, the depression and non-depression classes were separated and split while maintaining the same proportional distribution within each class. This ensured that each subset had an equal representation of both classes [15].

After the split, each subset was shuffled again to guarantee a random distribution within each set [13], some considerations regarding the oversampling with replacement (upsampling) process included identifying the majority class (non-depression). Consequently, oversampling with replacement was applied to the minority class (depression).

A random oversampling with replacement technique was applied to the depression class, which represented 31.90% of the total samples. This involved randomly selecting samples from the minority class and duplicating them until the number of examples matched that of the majority class (5,927 instances).

The original examples from the majority class (no depression) were combined with the duplicated examples from the minority class (depression) to form a new balanced dataset. To ensure a uniform distribution and prevent any order or pattern that could influence the model's learning process, a random shuffle of the dataset rows was performed.

Once the dataset was balanced, with a total of 11,854 samples (5,927 per class), it was divided into three subsets for the different stages of training and model evaluation.

3.4 Tokenization for Transformer Models

For the text to be processed by the Transformer model used, it must be converted into an appropriate semantic numerical representation through tokenization. In this project, three tokenizers were employed, each tokenizer had to adhere to a maximum sequence length of 128 tokens. This restriction is crucial as it limits sequence length, helping to reduce computational time and memory usage during training.

Additionally, batch size consistency was considered, as Transformer models require input dimensions to be uniform.

In truncation, sequences exceeding 128 tokens are cut to fit the maximum length, preventing excessively long data from affecting the model's performance. In padding, sequences shorter than 128 tokens are filled with special tokens to reach the required length, ensuring that all inputs have the same dimensions.

This tokenization and preparation process is crucial for transformer-based models to efficiently process the text and extract relevant features for the classification task.

3.5 Fine Tuning Model Based on Transformers

This stage of the proposed methodology focuses on applying fine-tuning to transformer-based architectures such as BERT, RoBERTa, and DeBERTa, combined with the use of contextual embeddings, to analyze user-generated texts and detect potential signs of depression. These representations enable the model to capture complex semantic nuances and deep contextual relationships within natural language.

The system was designed as a binary classifier (depression / non-depression) and is intended for use in psychological analysis contexts based on textual data provided by individuals.

Due to linguistic variability among users in terms of style, tone, and syntactic structures, the model presents significant challenges during training. To address these challenges and improve training effectiveness, the following strategy was applied.

The fine-tuning process was conducted on a manually annotated dataset, with specific adjustments to key training hyperparameters, such

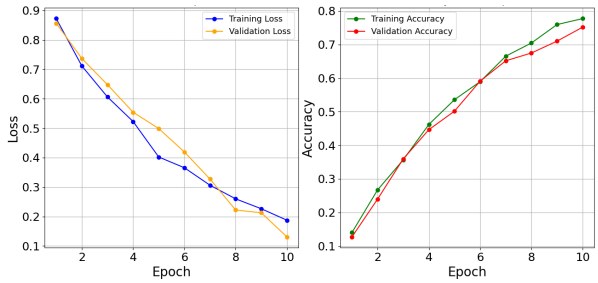

Fig. 2. Loss and accuracy graphs of the BERT model observed over 10 epochs

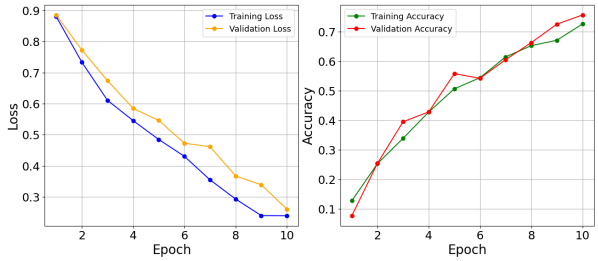

Fig. 3. Loss and accuracy graphs of the RoBERTa model observed over 10 epochs

as the number of epochs, learning rate, and batch size, in order to optimize the balance between accuracy and computational efficiency.

In particular, a set of experiments (grid search) showed that the most significant results were obtained using a learning rate of 2e-5, a batch size of 8 units, and a weight decay of 0.0001, training the model for 10 epochs.

The Adam optimizer was selected for its suitability with transformer-based models, particularly due to its incorporation of weight decay, which contributes to improved generalization. Furthermore, early stopping and checkpointing were applied to save the best performing model based on the F1 score metric, thus preventing overfitting and ensuring that the final model achieved optimal performance in the validation set.

4 Experiments and Results

To ensure the stability and reproducibility of the results, multiple validation subsets were processed through cross-validation. For each fold (k = 10), the model weights were randomly initialized, and training was conducted from scratch. Running multiple experiments with different initializations allows for evaluating the robustness and consistency of the model. This practice is widely recommended in the field of machine learning to mitigate the impact of randomness in initialization and to provide more reliable performance estimates [10]. The implementation of ten epochs is supported by previous research in the field of depression detection using BERT-based models [4].

For text processing, the input sequence length was set to a maximum of 128 tokens, after evaluating performance with alternative configurations of 64 and 256 tokens. This length was selected as it offered the best trade-off between model accuracy and computational efficiency.

The validation process was carried out using the following datasets, the training set included a total of 8,298 instances and was used to optimize the model's weights. During the exploratory data analysis phase, as described in Sections 3.1, 3.2, and 3.3, a balanced distribution of labels was observed, along with consistent lexical patterns and highly representative terms, which provided valuable discriminative information for the modeling process. The test dataset comprised 1896, validation dataset 1,660 instances and was used to monitor performance after each epoch, enabling the adjustment of hyperparameters as needed to improve generalization. Training, test and validation performance for BERT, RoBERTa, and DeBERTa models was monitored over 10 epochs.

Epochs was limited due to computational resource constraints, which influenced considered in the experimental design. Despite this limitation, the loss and accuracy curves provided meaningful insights into the learning behavior and generalization capacity of each model.

In Figure 2, BERT demonstrated convergence within the limited epochs, with steadily decreasing training and validation loss and stabilization of accuracy by the final epochs. It achieved an average acurracy 0.76 indicating effective pattern recognition under constrained training conditions.

RoBERTa and DeBERTa also showed promising results (Figure 3 and Figure 4, respectively),

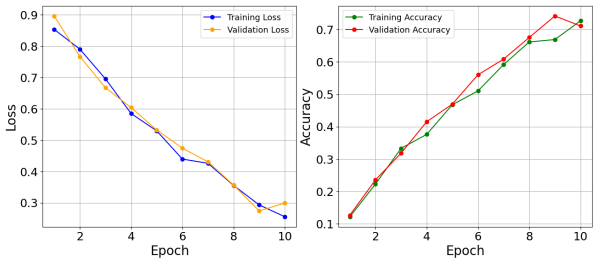

Fig. 4. Loss and accuracy graphs of the DeBERTa model observed over 10 epochs

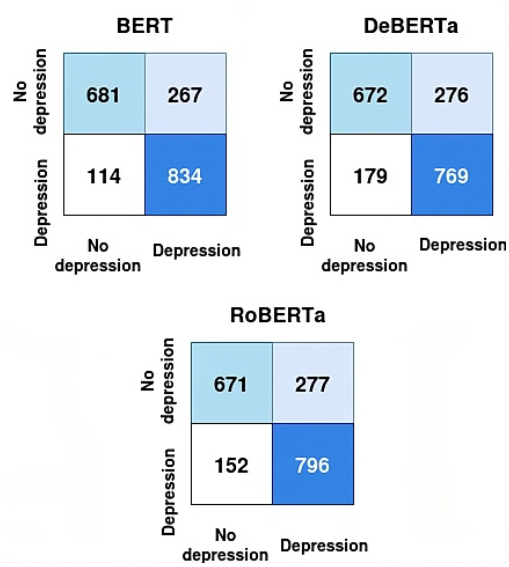

Fig. 5. Results confusion matrices of fitted models for the depression detection task

though DeBERTa exhibited slightly more stable validation loss across epochs, suggesting better generalization. However, BERT stood out as the most effective model given the convergence behavior and overall performance within the limited training window.

4.1 Confusion Matrices

The confusion matrices shown in Figure 5 provide information on the results obtained from the experiments carried out. Below, we analyze some instances to determine which pattern was possibly detected with the DAC-WOZ instances.

The Bert model with fine tuning correctly classified 681 examples as non-depression (true negatives) and 834 examples as depression (true positives). However, 267 examples were incorrectly classified as depression when they actually belonged to the non-depression class (false positives), and 114 examples from the depression class were misclassified as non-depression (false negatives). The low number of false negatives indicates that the model is effective at identifying cases of depression, although the relatively high number of false positives suggests a conservative tendency to classify toward the depression class.

DeBERTa model with fine tuning correctly classified 672 examples as non-depression and 769 examples as depression. Compared to BERT, this model exhibited a higher number of errors, with 276 examples incorrectly classified as depression and 179 depression cases misclassified as non-depression. Although DeBERTa showed a slight improvement in classifying the non-depression class, it produced a higher number of false negatives, which can be critical in the context of depression detection, where minimizing undetected cases is preferable.

The confusion matrix for the RoBERTa model with fine tuning indicates that it correctly classified 671 examples as non-depression and 796 examples as depression. It made 277 errors by classifying non-depression examples as depression (false positives) and 152 errors by classifying depression examples as non-depression (false negatives). RoBERTa demonstrated a reasonable balance between false positives and false negatives, although its overall performance is intermediate compared to BERT and DeBERTa.

Below are six textual examples that were input into the BERT depression detection model, along with the corresponding classification results (Table 3). The objective is to analyze their linguistic and structural characteristics in detail. This analysis allows for the identification of recurring patterns in the model's responses and, consequently, enables a qualitative interpretation of the confusion matrices. In this way, a more comprehensive and contextualized explanation of the classification errors is provided.

Table 3. Example sentences for the detection of tendency to depression

Word length	User input simulation	Status emotional	Time Response	Result prediction
7	Feeling drained bad and overwhelmed is terrible	Negative	98 ms	Tends to Depression
6	Today I achieved a great milestone	Positive	65 ms	Does not tend to Depression
11	I feel exhausted and overwhelmed all is terrible today really bad	Negative	100 ms	Tends to Depression
12	I feel motivated and excited after achieving my goals today	Positive	112 ms	Does not tend to Depression
19	I feel exhausted and hopeless today because nothing seems to be going right, and everything feels overwhelming and draining	Negative	120 ms	Tends to Depression
20	I feel incredibly happy and grateful today because everything is going well, and I am surrounded by wonderful people	Positive	121 ms	Does not tend to Depression

Table 4. Comparison of experiments of the BERT model

Training time	F1 score	ROC-AUC	Accuracy
50.0 m 8.06 s	0.771463	0.821695	0.782187
50.0 m 7.35 s	0.781325	0.840019	0.772614
50.0 m 16.77 s	0.787597	0.839821	0.734121
50.0 m 3.01 s	0.781043	0.625929	0.779880
50.0 m 4.92 s	0.694217	0.700303	0.512009
50.0 m 8.62 s	0.771226	0.834832	0.769167
50.0 m 19.11 s	0.785103	0.829915	0.756615
50.0 m 19.04 s	**0.799767**	**0.847312**	**0.781445**
50.0 m 21.26 s	0.772146	0.829342	0.756710
50.0 m 14.53 s	0.769627	0.820612	0.749185

Some responses in the DAIC-WOZ dataset are brief or unexpressive (e.g., ”yes”, ”I don’t know”), which limits the ability of models such as BERT, RoBERTa, and DeBERTa to extract meaningful semantic signals. An analysis of the dataset’s phrases reveals a clear polarity in certain expressions, with both positive and negative tendencies, often exaggerated in intensity. As a result, these models lack explicit emotional understanding: while they can identify linguistic patterns, they may not effectively capture the underlying emotional intent when it is not clearly expressed. Furthermore, the high computational cost of processing may have restricted the available training time, hindering the models’ ability to effectively learn relevant patterns from the dataset’s text samples.

4.2 Evaluation Metrics

This section analyzes the performance of the trained (fine-tuned) BERT, RoBERTa, and DeBERTa models for the text classification task, focusing on key metrics such as F1 score, ROC-AUC, and accuracy through cross-validation experiments, the highlighted experiment indicates better results obtained for that specific model, for each model, as it allows to evaluate both accuracy and stability of the models in a reliable way, the results are presented in Tables 4, 5 and 6, respectively. In addition, the computational cost and stability of the models are discussed in relation to the experiments performed. Their performance

is also compared with models from previous works reported in the literature.

The BERT model demonstrated the best overall performance among the evaluated architectures, with an average F1-Score $\approx 0.7691 \pm 0.0306$ (Table 4). This result indicates its ability to balance accuracy and sensitivity in the depression detection task. In addition, their average ROC-AUC was higher than 0.82 in some experiments, reflecting their ability to distinguish between the classes. On the other hand, as can be seen in the Accuracy column of Table 4, its average accuracy remained stable at values close to 0.76, consolidating it as the most consistent and reliable model. In terms of computational efficiency, BERT completed each experiment in approximately 50 minutes.

RoBERTa, although it showed competitive results, presented greater variability among experiments. Its average F1-Score $\approx 0.7219 \pm 0.0316$ (Table 5) suggests consistency problems, especially in experiments where the F1-Score remained at 0.6667. This instability may be due to a high sensitivity to weight initialization. Its average ROC-AUC, although adequate (0.61-0.78), was lower than that of BERT, and in specific experiments, fell below 0.61, indicating problems in class separation. RoBERTa required an average of 52 minutes per experiment, positioning it as a less efficient alternative to BERT.

Table 5. Comparison of experiments of the RoBERTa model

Training time	F1 score	ROC-AUC	Accuracy
52.0 m 49.17 s	0.656697	0.617193	0.662379
52.0 m 21.95 s	0.697147	0.671899	0.623129
52.0 m 9.23 s	0.741201	0.701591	0.746612
52.0 m 46.17 s	0.736893	0.711474	0.735721
52.0 m 18.35 s	0.721407	0.730063	0.704819
52.0 m 41.47 s	0.716667	0.629046	0.650110
54.0 m 11.91 s	**0.750914**	**0.781249**	**0.750011**
52.0 m 50.09 s	0.761547	0.684160	0.768670
52.0 m 51.42 s	0.666667	0.676141	0.651300
52.0 m 47.14 s	0.686667	0.712814	0.612430

DeBERTa presented an average F1-Score $\approx 0.7421 \pm 0.0412$ (Table 6), slightly outperforming RoBERTa but with a higher standard deviation. While some experiments achieved competitive metrics ($F1 - Score > 0.75$ and ROC-AUC > 0.77), others did not converged adequately experiments, evidencing their F1-Score sensitivity at 0.69 in multiple initial parameters and training configurations. Its average training time $\sim$82 minutes was significantly longer than that of the other models, increasing its computational cost.

Table 6. Comparison of experiments of the DeBERTa model

Training time	F1 score	ROC-AUC	Accuracy
77.0 m 15.41 s	0.762796	0.790917	0.733734
83.0 m 9.46 s	**0.785142**	**0.809713**	**0.748134**
81.0 m 55.93 s	0.748032	0.785600	0.730120
80.0 m 10.28 s	0.654467	0.539146	0.500000
79.0 m 12.98 s	0.733967	0.724199	0.689759
80.0 m 4.12 s	0.696467	0.535043	0.510269
81.0 m 13.42 s	0.696127	0.515090	0.501240
84.0 m 18.14 s	0.759149	0.819615	0.723916
82.0 m 10.46 s	0.759587	0.790623	0.751506
81.0 m 10.36 s	0.690267	0.535812	0.501250

The execution time is reported as evidence of the complexity involved in fitting more robust models to the data and the specific task, even when using rented high-performance hardware. The process was executed on a pay-per-use computing cluster, where access and performance depend on availability and the shared load among users.

The ROC-AUC analysis for the models evaluated reflects key differences in their ability to distinguish between classes. BERT stands out with consistently high values (> 0.82), indicating an excellent balance between false positives and true positives.

RoBERTa and DeBERTa showed lower values and considerable variation, evidencing lower performance in accurately identifying classes. Accuracy confirms the advantage of BERT, with stable values above 0.76, while RoBERTa and DeBERTa show a greater dependence on the accuracy of BERT. DeBERTa shows a greater dependence on the specific conditions of each experiment.

BERT consolidates as the most suitable model for the depression detection task, standing out for its stability, efficiency, and ability to outperform both the evaluated alternatives and the hybrid models reported in the literature. RoBERTa and DeBERTa, although promising, require additional adjustments in their configurations to improve their consistency and optimize their performance.

4.3 Comparison with related work

In the task of detecting depression from text, using the F1-score is particularly appropriate due to the natural class imbalance—there are typically more examples of individuals without symptoms than those with depression. In such scenarios, metrics like accuracy can provide a misleading view of model performance, as high overall accuracy may mask a poor ability to identify actual cases of depression (false negatives).

The F1-score, by combining precision (how reliable positive predictions are) and recall (how well actual positive cases are identified), offers a more balanced and representative measure of model performance. This is especially important in sensitive tasks like this, where failing to detect a case can have serious implications for timely intervention.

Comparisons with previous work using the same dataset are presented in Table 7, for example, the BERT + RoBERTa hybrid model with layers of attention and LSTM achieved an average F1-Score of 0.62 ± 0.12, while RoBERTa with BiLSTM achieved an F1-Score of 0.69. These values are lower than those obtained with the fine tuning architectures implemented in this work, highlighting the effectiveness of BERT, RoBERTa and DeBERTa when modifying the preprocessing performed to the DAIC-WOZ dataset.

It is worth noting, as mentioned at the beginning, that the DAIC-WOZ dataset is multimodal. However, this research focuses exclusively on the textual modality, in contrast, some related studies also incorporate audio data, as is the case with the work by Vandana presented in Table 7. Therefore, the comparison of F1-scores remains relevant, provided that the particular implementation conditions of our work are taken into account in relation to other approaches that use multiple modalities such as video, text, and audio.

Table 7. A comparison between the best F1 scores obtained by the models proposed in this study and those reported in the relevant literature.

Architecture	Average and standard deviation of F1 Score
BERT Fine tuning	0.774423088 ± 0.035982689
RoBERTa Fine tuning	0.724512882 ± 0.041573305
DeBERTa Fine tuning	0.719610796 ± 0.04699342
BERT + RoBERTa + attention layer + LSTM layer [14]	0.62000000 ± 0.12000000
RoBERTa + BiLSTM [22]	0.69000000
Text CNN [10]	0.60000000

5 Conclusion and Future Work

Notably, the implementation of the fine-tuning model was based on robust and previously validated natural language processing architectures, such as BERT, RoBERTa, and DeBERTa, due to the limited availability of clinically validated and professionally annotated datasets in the field of mental health. This lack of reliable data poses a significant challenge in the development of predictive models for mental health, as using unvalidated information can lead to inaccurate outcomes or inappropriate clinical interpretations.

Moreover, unlike other areas of machine learning, generating synthetic data or applying conventional data augmentation techniques, such as lexical substitution or automatic paraphrasing, is not feasible due to the sensitive, ethical, and clinical nature of content related to mental disorders. Manipulating this type of data may introduce serious biases, trivialize genuine symptoms, or result in unrealistic examples that fail to capture the complexity of language associated with depressive states. Furthermore, any misclassification by a model trained on poorly controlled data could have harmful consequences, particularly if used as a decision-support tool in clinical contexts.

Therefore, the decision to use pretrained models on curated corpora, combined with careful

adaptation through fine-tuning on datasets meticulously selected and annotated by domain experts, represents a responsible and methodologically sound approach. This strategy aims to balance the potential of NLP and large language models in the mental health domain with the imperative to uphold high ethical, scientific, and clinical standards in the development and deployment of computational models for human language analysis in sensitive contexts.

Future work includes improving the system's accuracy, accessibility, and scalability, with the goal of strengthening its potential as an effective tool for the robust detection of depressive symptoms. To achieve this, it is proposed to explore advanced NLP techniques, such as the incorporation of multilingual models, the use of deeper contextual representations, the integration of sentiment and emotion analysis, and the application of self-supervised learning methods. These enhancements would not only increase the system's accuracy across diverse contexts, but also facilitate its deployment on platforms accessible to a broader audience and adaptable to varying workloads and data volumes.

References

1. **Anderson, T., Sarkar, S., Kelley, R. (2024).** Analyzing public sentiment on sustainability: A comprehensive review and application of sentiment analysis techniques. Natural Language Processing Journal, pp. 100097.
2. **Danner, M., Hadzic, B., Gerhardt, S., Ludwig, S., Uslu, I., Shao, P., Weber, T., Shiban, Y., Ratsch, M. (2023).** Advancing mental health diagnostics: Gpt-based method for depression detection. 2023 62nd Annual Conference of the Society of Instrument and Control Engineers (SICE), IEEE, pp. 1290–1296.
3. **DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., et al. (2014).** Simsensei kiosk: A virtual human interviewer for healthcare decision support. Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems, pp. 1061–1068.
4. **Gratch, J., Artstein, R., Lucas, G. M., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., et al. (2014).** The distress analysis interview corpus of human and computer interviews. LREC, Reykjavik, Vol. 14, pp. 3123–3128.
5. **Hadzic, B., Mohammed, P., Danner, M., Ohse, J., Zhang, Y., Shiban, Y., Rätsch, M. (2024).** Enhancing early depression detection with ai: a comparative use of nlp models. SICE journal of control, measurement, and system integration, Vol. 17, No. 1, pp. 135–143.
6. **Huang, R., Yi, S., Chen, J., Chan, K. Y., Chan, J. W. Y., Chan, N. Y., Li, S. X., Wing, Y. K., Li, T. M. H. (2024).** Exploring the role of first-person singular pronouns in detecting suicidal ideation: A machine learning analysis of clinical transcripts. Behavioral Sciences, Vol. 14, No. 3, pp. 225.
7. **Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B., Berry, J. T., Mokdad, A. H. (2009).** The phq-8 as a measure of current depression in the general population. Journal of affective disorders, Vol. 114, No. 1-3, pp. 163–173.
8. **Li, N., Wu, D. D. (2010).** Using text mining and sentiment analysis for online forums hotspot detection and forecast. Decision support systems, Vol. 48, No. 2, pp. 354–368.
9. **Mao, H., Han, Q. (2025).** Applications of transformer-based language models for depression detection: A scoping review. Journal of Integrated Social Sciences and Humanities, pp. 1–8.
10. **Marriwala, N., Chaudhary, D., et al. (2023).** A hybrid model for depression detection using deep learning. Measurement: Sensors, Vol. 25, pp. 100587.
11. **Medina-Mora, M. E., Orozco, R., Rafful, C., Cordero, M., Bishai, J., Ferrari, A., Santomauro, D., Benjet, C., Borges, G., Mantilla-Herrera, A. M. (2023).** Los trastornos

mentales en méxico 1990-2021. resultados del estudio global burden of disease 2021. Gaceta médica de México, Vol. 159, No. 6, pp. 527–538.

12. **Nushida, T., Kang, X., Matsumoto, K., Yoshida, M., Zhou, J. (2025).** An automated depression diagnosis system utilizing a knowledge base created with gpt. 2025 IEEE 17th International Conference on Computer Research and Development (ICCRD), IEEE, pp. 329–333.
13. **Prodregosa, F., et al. (2011).** Scikit-learn: Machine learning in python. journal of machine learning research. vol, Vol. 12, pp. 2825–2830.
14. **Senn, S., Tlachac, M., Flores, R., Rundensteiner, E. (2022).** Ensembles of bert for depression classification. 2022 44th annual international conference of the IEEE engineering in Medicine & Biology Society (EMBC), IEEE, pp. 4691–4694.
15. **Shalev-Shwartz, S., Ben-David, S. (2014).** Understanding machine learning: From theory to algorithms. Cambridge university press.
16. **Swain, V. D., Ye, J., Ramesh, S. K., Mondal, A., Abowd, G. D., De Choudhury, M., et al. (2024).** Leveraging social media to predict covid-19–induced disruptions to mental well-being among university students: Modeling study. JMIR Formative Research, Vol. 8, No. 1, pp. e52316.
17. **Teng, S., Zhang, T., D'Alfonso, S., Kostakos, V. (2024).** Predicting affective states from screen text sentiment. Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 384–390.
18. **University of Southern California, Institute for Creative Technologies (2022).** Daic-woz: Distress analysis interview corpus - wizard of oz. `https://dcapswoz.ict.usc.edu/`. Accessed: 2025-05-30.
19. **University of Southern California, Institute for Creative Technologies (2022).** Daic-woz: Distress analysis interview corpus - wizard of oz pdf. `https://surli.cc/qrdruh`. Accessed: 2025-07-10.
20. **University of Southern California, Institute for Creative Technologies (2022).** Daic-woz end-user license agreement. `https://dcapswoz.ict.usc.edu/daic-woz-database-download/`. Accessed: 2025-07-10.
21. **Wurtz, H. M., Willen, S. S., Mason, K. A. (2022).** Introduction: Journaling and mental health during covid-19: Insights from the pandemic journaling project. SSM-Mental Health, Vol. 2, pp. 100141.
22. **Zhang, J., Guo, Y. (2024).** Multilevel depression status detection based on fine-grained prompt learning. Pattern Recognition Letters, Vol. 178, pp. 167–173.

Article received on 30/05/2025; accepted on 05/09/2025.
**Corresponding author is Lauro Reyes-Cocoletzi.*