Detecting anxiety and depression in dialogues: a multi-label and explainable approach

Francisco de Arriba-Pérez[1,†], Silvia García-Méndez[1,*,†]

[1]*Information Technologies Group, atlanTTic, University of Vigo, Vigo, Spain*

Abstract

Anxiety and depression are the most common mental health issues worldwide, affecting a non-negligible part of the population. Accordingly, stakeholders, including governments' health systems, are developing new strategies to promote early detection and prevention from a holistic perspective (*i.e.*, addressing several disorders simultaneously). In this work, an entirely novel system for the multi-label classification of anxiety and depression is proposed. The input data consists of dialogues from user interactions with an assistant chatbot. Another relevant contribution lies in using Large Language Models (LLMs) for feature extraction, provided the complexity and variability of language. The combination of LLMs, given their high capability for language understanding, and Machine Learning (ML) models, provided their contextual knowledge about the classification problem thanks to the labeled data, constitute a promising approach towards mental health assessment. To promote the solution's trustworthiness, reliability, and accountability, explainability descriptions of the model's decision are provided in a graphical dashboard. Experimental results on a real dataset attain 90 % accuracy, improving those in the prior literature. The ultimate objective is to contribute in an accessible and scalable way before formal treatment occurs in the healthcare systems.

1. Introduction

More than 55 million people in the United States suffer from mental illness as indicated by the National Institutes of Health (NIH, 2023)[1]. More in detail, the most common mental conditions are anxiety (19.1 %) and major depression (8.3 %). At the global level, 4 % of the population is affected by anxiety disorder. At the same time, 280 million people worldwide suffer from depression, as stated by the World Health Organization (WHO, 2023)[2]. However, only 25 % of people suffering from anxiety receive treatment. A recent report by Forbes[3] completes this information and indicates that 50 % of people affected by depression go undiagnosed in the primary care system.

In this regard, it should be noted that traditional screening methods (*i.e.*, those that rely on subjective and time-consuming interviews composed of binary questions for patients and their families) face several issues [1]. Among them, the unreliability of self-reported diagnoses due to bias introduced by subjectivity, intentional concealment, and even inconvenience of the number of questions must be considered, resulting in the latter low rates of diagnosis and interventions [2]. Another concern is stigma, which prevents treatment seeking and ignorance of the condition [1]. Representative examples of these self-reporting methodologies are the Beck Depression Inventory (BDI), the General Health

*Corresponding author.

† These authors contributed equally.

✉ farriba@gti.uvigo.es (F. d. Arriba-Pérez); sgarcia@gti.uvigo.es (S. García-Méndez)

🌐 https://franciscodearribaperez.ddnsfree.com (F. d. Arriba-Pérez); https://silviagarciamendez.ddnsfree.com (S. García-Méndez)

🆔 0000-0002-1140-679X (F. d. Arriba-Pérez); 0000-0003-0533-1303 (S. García-Méndez)

[1]Available at https://www.usa.edu/blog/mental-health-statistics, October 2024.

[2]Available at https://www.who.int/news-room/fact-sheets/detail/depression, https://www.who.int/news-room/fact-sheets/detail/anxiety-disorders, October 2024.

[3]Available at https://www.forbes.com/health/mind/depression-statistics, October 2024.

Questionnaire (GHQ), the Hamilton Rating Scale for Depression (HRSD), and the Patient Health Questionnaire (PHQ). Similar to the GHQ and the PHQ, the Depression, Anxiety, and Stress Scale (DASS) combines the questionnaires of each factor [3]. Consideration should also be given to the popular Diagnostic and Statistical Manual of Mental Disorders, fifth edition (DSM-5) published by the American Psychiatric Association.

Provided the severe consequences of anxiety and depression that even increase the risk of suicide [4], early detection and timely diagnoses are critical. In this regard, language can be a good predictor of mood disorders [5]. More in detail, how users engage in a conversation and express themselves is a strong indicator of their mental health state. Accordingly, the arrival of Large Language Models (LLMs, *e.g.*, GPT-4[4], PaLM[5], and Alpaca[6]) has contributed significantly to health-related topics thanks to their context-learning capabilities, mostly in generative tasks. Specifically, the literature has reported promising performance of these models in three relevant scenarios: (*i*) language comprehension, (*ii*) text generation, and (*iii*) knowledge inference [6]. Moreover, the potential of these models to leverage large volumes of online data is of great importance for both diagnosis and treatment [7].

Consequently, several pre-trained language models (PLMs) and LLMs have been deployed for addressing health issues like mental disorders. It is the case of LLMental [8], MentalBERT [9] and MentalLLaMA [10]. Besides, PsychBERT [11] is fine-tuned to detect language patterns in behavioral health, mental health, psychiatry, and psychology texts. However, as indicated by previous works [1], their performance in specific classification problems with task-specific data like anxiety and depression is still immature when used as final solutions (*i.e.*, in zero-shot/few-shot learning or with limited fine-tuning). This is due to the poor detection of nuances and task-specific patterns essential for accurate detection. Similarly, the PLMs exhibit limited generalization and low multitask robustness [12, 13]. Another key limitation is their low interpretability, which prevents their practical use beyond academic research [10].

Summing up, using PLMs or directly LLMs in zero-shot settings is the prominent approach [14]. Regarding the experimental data, most researchers use social media [15]. In recent years, there has even been an increasing interest in detecting mental health states with tracking devices, which results from the growing importance that modern society places on mental well-being [16]. Regardless of the approach, the aspect in which most of the researchers meet is the necessity to provide interpretable results along with explainable descriptions of the rationale of the machine-based solutions, of uttermost importance in the healthcare field provided their direct impact in the decision-making of clinicians and thus, the patient's well-being. In this regard, eXplainable Artificial Intelligence (XAI) comprises post hoc and self-explanatory techniques. While the former posthoc alternatives aim to explain the prediction of black-box classification models like the popular explanatory model-agnostic tools (*i.e.* those that combine local linear and random models, like LIME and SHAP [17] to approximate feature importance weights with regression and game theory), the self-explanatory approach relies on intrinsically interpretable models that can provide explanations along with the predictions [14]. However, feature importance methods like LIME and SHAP only provide the weight of the selected features without considering the interactions among the features and are low intuitive for end users [18]. In this regard, a major regulatory milestone in the AI field was materialized with the Artificial Intelligence Act (AIA). The final text pays particular attention to interpretability, the right of end users to receive clear explanations, and the disclosure of the use of AI in human interactions [19].

Given the safety-critical nature of these conditions, our solution must provide high accuracy and explainability to promote trust among the end users and professionals. Accordingly, we combine the traditional Machine Learning (ML) models (which can offer higher accuracy but lack explainability) operating in a multi-label setting with LLMs (which are intrinsically explicable but lack specific downstream knowledge). Note that in our approach, LLMs are leveraged to extract users' expert features related to anxiety and depression by detecting linguistic patterns and language usage, taking advantage of their understanding capabilities. Specifically, relying on LLMs solely as part of the feature engineering

---

module to extract user-level knowledge; we tackle the hallucination problem, that is, those predictions that, even seem correct, present underlying misconceptions due to the absence of a comprehensive understanding of the problem and expert data. Furthermore, accurate diagnosis requires formal clinical knowledge [5]. Consequently, in this work, we understand the necessity of leveraging formal medical knowledge into machine-based solutions (*i.e.*, by incorporating transparent assessment based on official methods, scales, and standards). Hence, we used formal clinical scales for anxiety and depression to label the experimental data. We also acknowledge the limitations of using social media data. Thus, we exploit free dialogues with a conversational assistant. Compared to free dialogues, clinical questionnaires limit the users' ability to express their state freely. Our ultimate objective is to perform an on-demand and scalable assessment of anxiety and depression before formal clinical screening in the healthcare systems. Note that intentional concealment is reduced in our study since the tests are embedded in the dialogues, and the questions are adapted accordingly.

The remainder of this manuscript is divided into the following sections. Section 2 summarizes the key prior works on anxiety and depression detection using PLMs and LLMs, paying particular attention to multi-label approaches and those that provide explainability. Section 3 details our system architecture, while Section 4 shows the results obtained with our methodology and compares them with other works in the state of the art. Finally, Section 5 does the main conclusions of this work and proposes future research.

2. Literature review

Traditional ML, deep learning, and Natural Language Processing (NLP) techniques have been used in the literature for mental health assessment [3]. The most recent works involve word embedding with transformed-based models (also known as PLMs) to take advantage of contextual data [20, 21]. However, scant research is available in state of art related to using LLMs like ChatGPT underlying models; most use them as final classification solutions with limited prompt engineering or fine-tuning.

Regarding the detection of anxiety and depression, many researchers apply a multitask learning perspective (*i.e.*, defining a primary and an auxiliary task), *e.g.*, emotion inference. This is due to the availability of experimental labeled data in terms of emotional content [22]. These works sustain that stressed users are more likely to express negative emotions (*e.g.*, anger, fear, and sadness) rather than positive ones (*e.g.*, happiness). This is the case in many works. In this regard, Qureshi et al. [23] defined emotion classification as the second task, which follows depression as the main task, similar to what Ghosh et al. [24] proposed. Moreover, the solution developed by Turcan et al. [25] is another representative example. Notably, the authors applied this approach to stress detection. They explored single-task models that operate similarly to BERT and multitask learning with a fine-tuned BERT model on emotion detection and stress labels. Finally, they exploited LIME for interpretability. Although we agree with the strong relation between emotion load and mental health state, we believe that relying mainly on emotion detection to assess anxiety, depression, or stress may lead to false positive results. Thus, we incorporated this knowledge into the engineered features, using anxiety and depression-labeled data as main tasks jointly in a multi-label setting.

Moreover, Ghosh et al. [26] proposed a multitasking framework (not based on ML models) for depression detection, sentiment classification, and emotion recognition. Even if slightly related to our research, the promising results obtained prove the appropriateness of addressing anxiety and depression simultaneously, given the strong link between both mental health conditions. Conversely, Sarkar et al. [27] developed a multitask learning solution with a data-sharing mechanism, providing the relation between anxiety and depression. The authors used word embedding models like BERT for feature engineering to feed traditional ML models, similar to our work but without the advantage in terms of explainability that LLMs provide. Alike to the work by Sarkar et al. [27] is the more recent proposal by Park et al. [28]. Additionally, Ilias and Askounis [29] defined a multitask learning framework in which depression and stress detection are the main and auxiliary tasks, respectively, using social media data. Note that two datasets gathered and labeled in different conditions are used. The first proposed

approach encompasses a BERT-based layer shared for both tasks, primary and auxiliary, followed by separate BERT-based encoder layers. In contrast, the second approach derives from the first but exploits weighting layers by attention fusion networks. However, no hyperparameter tuning was performed due to limited access to computational resources. Explainability was not provided either.

Despite the strong relation between anxiety or stress and depression, few studies address the joint assessment of several conditions [29]. In this regard, Lee et al. [2] focused on geriatric (*i.e.*, experimental data from mild cognitive impairment patients) anxiety and depression detection by exploiting low-cost activity trackers. Regarding the multi-label classification approach followed, the authors applied the binary relevance method. That is, unlike in our work, they used two single-label classifiers for anxiety and depression, respectively, which is a more straightforward way of approaching the problem. However, accuracy may be compromised since the solution does not consider the correlation between labels. As in our work, they include questionnaire-based features from the geriatric anxiety inventory (GAI) and the geriatric depression scale (GDS). In addition, Park et al. [30] also integrated the DSM-5 diagnostic criteria into their predictive methodology, which is based on a variant of the BERT model. Similarly, de Souza et al. [31] proposed a stacking solution with two single-binary classifiers for anxiety, depression, and their comorbidity leveraging social media data. Note that the authors used SHAP for interpretability.

Some authors exploited the already mentioned PLMs. It is the case of Ahmed et al. [1] who proposed a transformer-based architecture for multi-class depression detection (*i.e.*, in severity levels: absent, mild, moderate, and severe). After text processing, different variants of the BERT model are used for classification. The final result is obtained following a voting approach. The authors applied LIME to provide interpretability to the solution. Ultimately, the proposed system was compared with ChatGPT (gpt-3.5-turbo model, non-fine-tuned), which attained poor performance. Related to our work, Chowdhury et al. [7] studied early depression detection from social media data using LLMs (*i.e.*, GPT-4), deep learning (*e.g.*, LSTM) and transformer models (*e.g.*, BERT). However, the authors' approach to explanability is to provide feature-level interpretability. More recently, Ilias et al. [32] developed a transformer-based solution for stress and depression detection from social media data. Extra-linguistic information is introduced to the BERT and MentalBERT models. However, the solution does not approach the detection task simultaneously, as in our work, which is much more challenging. Conversely, experiments were performed with datasets for binary classification of stress and depression, respectively, and a multi-class (*i.e.*, with different severity levels) depression dataset.

When it comes to the application of LLMs, Wang et al. [6] leveraged a fine-tuned version of ChatGPT to detect depression. To ensure accurate predictions, the authors proposed a knowledge-enhanced pre-training scheme with emotion analysis capabilities and human feedback. Moreover, Liu et al. [14] used ChatGPT for data collection along with manually created psychology data that feed BERT and RoBERTa models for depression detection. Regarding interpretability, SHAP was exploited. Similarly, Ohse et al. [33] investigated several PLMs and LLMs (*e.g.*, BERT, GPT-4, LLaMA) for depression assessment using clinical interviews as experimental data. The authors exploited the models following the zero-shot paradigm without fine-tuning or prompt engineering. Despite being a relevant study to endorse the applicability to the mental health field of these models, the authors did not exploit their full potential, as already mentioned with the lack of tuning and also regarding explainability. Furthermore, Wang et al. [34] proposed a solution that searches for depression-related texts from the BDI questionnaire. Then, LLMs are used to fill the latter survey using user data from social media to infer their mental state. Ultimately, Xu et al. [15] evaluated different LLMs (*e.g.*, Alapaca, LLaMA, GPT-4) for mental health classification (binary and multi-class prediction for stress, depression and suicide) from social media data exploiting prompt engineering. Note that this work differs from ours in the absence of the multi-label setting and explainable capabilities.

2.1. Research contributions

Table 1 shows the most closely related solutions to easily compare and assess our contributions. To the best of our knowledge, our work is the first to apply LLMs to extract users' expert features related to

anxiety and depression. By this means, we can detect linguistic patterns and language usage, using the comprehension capabilities of llms without sacrificing explainability. Moreover, another relevant contribution is combining traditional ml models in a multi-label setting, which can offer higher accuracy. Consideration should also be given to integrating formal clinical knowledge through standard tests used for data labeling. More in detail, experimental data consists of a free conversation between patients and a conversational chatbot, despite the popularity of social media data for anxiety and depression detection and the rigidity of self-reporting questionnaires. Ultimately, an explainability dashboard describes the most relevant data that leads to the classification decision and its confidence.

Summing up, the main contributions of the proposed solution for the field are:

- A multi-label framework able to predict jointly anxiety and depression.
- The use of llms to extract high-level reasoning features used to train the ml models.
- The explainability dashboard which promotes trust and makes the solution accountable and reliable.

Table 1
Comparison of related solutions.

Authorship	Approach	Data	Explainability
Lee et al. [2]	Multi-label ml	Tracker and clinical data	✗
Wang et al. [6]	llms	General and depression related text	✗
Ohse et al. [33]	plms/llms	Clinical interviews	✗
Xu et al. [15]	llms	Social media	✗
Our proposal	Multi-label ml + llms	Dialogues + clinical data	✓

3. Methodology

Figure 1 shows the scheme of our proposal. It is composed of (*i*) data acquisition to gather and filter data; (*ii*) feature engineering to generate new features using prompt engineering through an llm model and applying sliding window treatment; (*iii*) feature analysis and selection to remove no-relevant features. Moreover, the (*iv*) classification module evaluates our multi-label solution. Finally, (*v*) the explainability module generates a natural language explanation of the classifier predictions leveraging a prompt engineering template.

3.1. Data acquisition

The experimental dataset is composed of conversations with *Celia* chatbot[7]. This chatbot establishes an entertaining and engaging dialogue with end users, including fun facts about the conversation topics. Moreover, every 3 months, the chatbot uses the standard questionnaires presented in the Spanish versions of the Goldberg Anxiety and Depression Scales (gads) and the Yesavage Geriatric Depression Scale (ygds) to assess the cognitive state of the user. These questions are embedded during the conversation flow. The latter data is used as the label of the user (*i.e.*, absence or presence of anxiety and depression) for the supervised learning stage.

3.2. Feature engineering

The solution combines feature generation based on prompt engineering with a sliding window strategy to consider the history of past sessions[8]. Table 2 shows the features engineered that can verse on the

---

[7]Available at https://celiatecuida.com/en/home_en, October 2024.
[8]A session is a complete dialogue with the end user until they decide to stop the conversation.
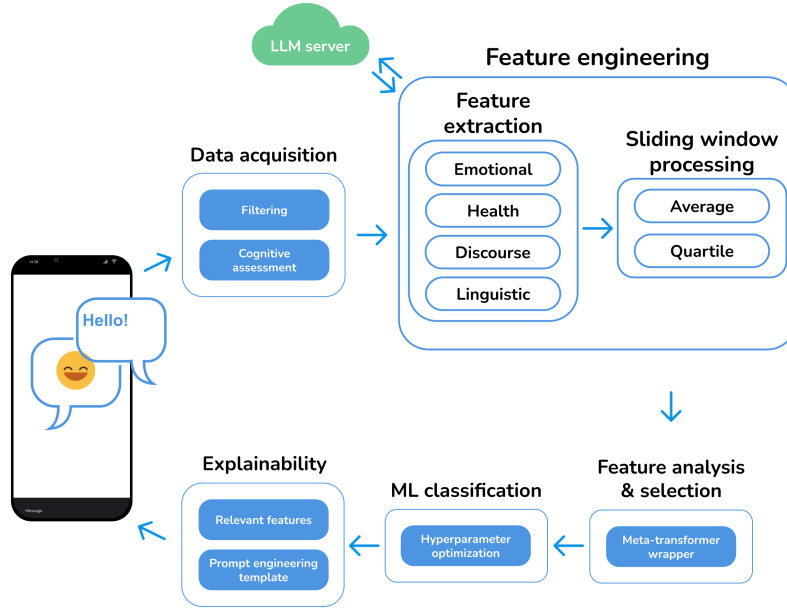
Figure 1: System scheme.

cognitive state of the end user (*i.e.*, their emotional well-being or health condition) or on the dialogue itself (*i.e.*, the discursive and linguistic characteristics of the conversation with the chatbot). These features are calculated using an LLM and prompt engineering, and their values range from 0.0 to 1.0[9].

Each generated feature is expanded with four new statistical features (average, and the three quartiles Q_1, Q_2, and Q_3). For this purpose, a sliding window is performed with the last 30 sessions (see Equation (1), where n is the number of sessions and $X[n]$ is the historical feature in the last n sessions. Consequently, $Y[n]$ is the ordered version of $X[n]$.

$$\forall n \in \{1...\infty\}$$

$$X[n] = \{x[0], \ldots, x[n]\}.$$
$$Y[n] = \{y_0[n], y_1[n], \ldots, y_{n-1}[n]\} \mid y_0[n] \leq y_1[n] \leq \ldots \leq y_{n-1}[n],$$
$$\text{where}; \forall x \in X[n], \ x \in Y[n].$$

$$avg[n] = \frac{1}{n} \sum_{i=0}^{n} y_i[n]$$
$$Q_1[n] = y_{\lfloor \frac{1}{4}n \rceil}[n]$$
$$Q_2[n] = y_{\lfloor \frac{2}{4}n \rceil}[n]$$
$$Q_3[n] = y_{\lfloor \frac{3}{4}n \rceil}[n]$$

(1)

3.3. Feature analysis & selection

In the cold-start step, the system uses $10\,\%$ of the samples to select the most relevant features. In this first phase, a selector based on a meta-transformer wrapper[10] is applied (*i.e.*, following a model-agnostic strategy). In order to select the most significant features, this transformer uses a tree ML classifier that

---

[9]Polarity takes integer values: 0 for negative, 1 for neutral, and 2 for positive.
[10]Available at https://scikit-learn.org/1.5/modules/feature_selection.html, October 2024.

Table 2
Features engineered for anxiety an depression detection.

Group	Category	ID	Name	Description
Cognitive state	Emotional	1	Insecurity	Level of insecurity or fear.
		2	Loneliness	Level of isolation.
		3	Negative emotion	Level of negative charge.
		4	Positive emotion	Level of positive charge.
		5	Sadness	Level of sorrow.
	Health	6	Anguish	Level of distress.
		7	Health issues	Level of health concerns.
Content	Discourse	8	Catastrophic terms	Level of use of disaster charge.
		9	Emphasized terms	Level of use of emphasis (positive or negative).
		10	Repeated concepts	Extent to which the user repeats previous content.
	Linguistic	11	Interjections	Level of use of interjections.
		12	Negative adverbs	Level of use of negative adverbs.
		13	Negatives terms	Level of use of negative terms.
		14	Polarity	Negative, neutral, positive.

calculates the Mean Decrease in Impurity (MDI)[11] of each feature. Ultimately, the features with a MDI lower than average are discarded.

3.4. Classification

In this study, we solve a multi-label binary classification problem. More in detail, this scenario differs from the binary class in the number of labels assigned. In the latter binary-class scenario, the classifier provides a single label between the two classes in the experimental data (*i.e.*, number of classes is 2, the number of resulting labels is 1). Since ours is a binary problem, we applied the Multi-class Transformation Strategy (MTS) [35]. Particularly, in our approach, we group the binary classes into four categories: (none_none), (none_depression), (anxiety_none), and (anxiety_depression). Note that the multi-label classification is a complex problem because the results may be partially correct, preventing the use of standard ML evaluation metrics. Instead, the metrics described below (micro and macro approach) are computed:

- **Exact match ratio** represents the proportion of predictions where both labels are correct (see equation (2), where A is the actual label, P is the predicted label, and n is the number of samples).

$$\frac{1}{n} \sum_{i=1}^{n} I(A_i = P_i) \tag{2}$$

- **Accuracy** is the percentage of correctly predicted labels over the total predicted and actual categories (see equation (3)).

$$\frac{1}{n} \sum_{i=1}^{n} \frac{|A_i \cap P_i|}{|A_i \cup P_i|} \tag{3}$$

- **Precision** is the percentage of correctly predicted labels over predicted labels (see equation (4)).

$$\frac{1}{n} \sum_{i=1}^{n} \frac{|A_i \cap P_i|}{|P_i|} \tag{4}$$

---

[11]Available at https://scikit-learn.org/1.5/auto_examples/inspection/plot_permutation_importance.html, October 2024.

- **Recall** is the percentage of correctly predicted labels over actual labels (see equation (5)).

$$\frac{1}{n}\sum_{i=1}^{n}\frac{|A_i \cap P_i|}{|A_i|} \tag{5}$$

- **Hamming Loss** (HL) calculates the incorrectly predicted labels. In our binary problem, it complements the accuracy (see equation (6)).

$$1 - Accuracy \tag{6}$$

Our solution exploits the Naive Bayes (NB), Decision Tree (DT), and Random Forest (RF) classifiers widely used in the literature to solve similar classification problems [2, 24, 27]. We analyze two different scenarios. Scenario 1 uses all user sessions to summarize the analysis, while scenario 2 evaluates the behavior of the system by reducing the number of samples and selecting 2 out of 3 entries.

3.5. Explainability

Our system explains the predictions obtained by leveraging an LLM with a prompt engineering template. This approach creates an explanation of the predicted majority category every 7 sessions. For this purpose and to limit the computational load of the explainability module, the most representative statistics of the features in Table 2 (*i.e.*, the average and the Q_2) are considered. Moreover, the conversations of the last two sessions are also sent to the model with the predicted category for interpretability purposes. In addition to promoting trust among end users and clinicians and the accountability and reliability of the solution, this information can be exploited to recommend formal assessment in the primary care health system or treatment to prevent anxiety and depression.

4. Results and discussion

This section explains the experimental dataset used and the results obtained. The analysis was conducted on a computer with the following specifications:

- **Operating System (OS)**: Ubuntu 18.04.2 LTS 64 bits
- **Processor**: IntelCore i9-10900K 2.80 GHz
- **RAM**: 96 GB DDR4
- **Disk storage**: 480 GB NVME + 500 GB SSD

4.1. Experimental data

The dataset[12] contains the complete conversations between voluntary users and the *Celia* chatbot from 16 May 2023 to 9 October 2024. Notably, it comprises 2186 user sessions, 32 users, and an average of 68 sessions per user. Moreover, each session comprises an average of 26 interactions and 157 words per session. Table 3 shows the distribution of sessions by categories. More in detail, most cases are concentrated in people without any pathology, and the presence of depression overlaps with anxiety, reducing the number of isolated cases of the former. This increases the difficulty of the classification problem, given the imbalance of experimental data, even more so in the multi-label scenario.

4.2. Data acquisition

To motivate a new conversation with *Celia*, the assistant sends notifications by email and shows reminders to the users. In this line, to detect the end of a session, the user must say goodbye, or this is automatically finished after 3 minutes of inactivity. Those sessions with 5 or fewer human interventions

---

[12]The experimental dataset is available on request from the authors.

Table 3
Experimental dataset of anxiety and depression.

Anxiety	Depression	Number of sessions
Absent	Absent	1287
Present	Absent	542
Absent	Present	29
Present	Present	328
Total		2186

are discarded to ensure that a significant amount of data enters the anxiety and depression detection system. In addition, as described in Section 3.1, every 3 months, the samples are re-labeled on anxiety and depression using the standard questionnaires GADS and YGDS.

4.3. Feature engineering

The features described in Table 2 are generated using the `gpt-4o-mini`[13] model. It can be accessed by sending requests to OpenAI API[14], using the prompt in Listing 1. Each request contains the text of the complete session and the temperature parameter set to 0. This removes the randomness and ensures that the model provides the same response to the same input content.

For each of the 14 features, the average and the three quartiles Q_1, Q_2, and Q_3 are calculated. In total, 56 features are used in this multi-label problem. Once calculated, the values are rounded to 2 digits, and those with the same values are discarded.

LISTING 1: Feature engineering using GPT-4o-mini.

```
This is a conversation between a bot and a person. It analyzes the interactions of
the person as a whole. Please indicate on a range from 0.0 to 1.0 the level of
insecurity, loneliness, negative emotion, positive emotion, sadness, anguish,
health issues, and the use of catastrophic terms, exaggerated terms, repeated
concepts, interjections, negative adverbs, negative terms, and polarity.
Following this JSON format. Do not add any textual explanation:
{insecurity:0,
loneliness:0,
negative_emotion:0,
positive_emotion:0,
sadness:0,
anguish:0,
health_issues:0,
catastrophic_terms:0,
exaggerated_terms:0,
repeated_concepts:0,
interjections:0,
negative_adverbs:0,
negatives_terms:0,
polarity:0 (0 negative, 1 neutral, 2 positive)}
<session>
```

4.4. Feature analysis & selection

Our approach uses the `SelectFromModel`[15] library of `scikit-learn` in combination with the RF classifier[16] to analyze and select the most relevant features. This analysis is performed using the 10 %

---

[13]Available at https://platform.openai.com/docs/models/gpt-4o-mini, October 2024.

[14]Available at https://openai.com/api, October 2024.

[15]Available at https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectFromModel.html, October 2024.

[16]Available at https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html, October 2024.

of the dataset. Specifically, 39 % of the original features are selected in scenario 1 and 2.

4.5. Classification

We evaluate our approach using the `scikit-learn` Python library. The NB[17], DT[18] and RF models are selected.

The hyperparameters of these classifiers are optimized using the `GridSearchCV`[19] method. We use a 10-fold cross-validation to evaluate the configuration that offers the best accuracy value over the 10 % of the experimental dataset. Listings 2, 3 and 4 contain the parameter sets for NB, DT, and RF, respectively. The values selected for scenarios 1 and 2 are shown below.

Scenario 1:

- NB: var_smoothing = 1e-05
- DT: splitter = best, max_features = None, max_depth = 100, min_samples_split = 0.001, min_samples_leaf = 0.001, criterion = entropy
- RF: n_estimators = 200, max_features = sqrt, max_depth = 10, min_samples_split = 2, min_samples_leaf = 1, criterion = gini

Scenario 2:

- NB: var_smoothing = 1e-05
- DT: splitter = random, max_features = sqrt, max_depth = 100, min_samples_split = 0.001, min_samples_leaf = 0.001, criterion = entropy
- RF: n_estimators = 100, max_features = None, max_depth = 5, min_samples_split = 2, min_samples_leaf = 1, criterion = entropy

LISTING 2: NB hyperparameter configuration.

```
var_smoothing : [1e-9, 1e-5, 1e-1]
```

LISTING 3: DT hyperparameter configuration.

```
splitter : [best, random],
max_features : [None, sqrt, log2],
max_depth : [1, 100, None],
min_samples_split : [0.001, 0.1, 1],
min_samples_leaf : [0.001, 0.1, 1],
criterion : [gini, entropy]
```

LISTING 4: RF hyperparameter configuration.

```
n_estimators : [100,150,200],
max_features : [sqrt, log2, None],
max_depth : [5, 10, 100, None],
min_samples_split : [2, 5, 10],
min_samples_leaf : [1, 2, 5],
criterion : [gini, entropy]
```

Table 4 shows the classification results obtained. Note that the three alternative models attain promising results, RF the one that attained the best values, all above 80 % regardless of the scenario. Moreover, the accuracy value is close to 90 % (*i.e.*, 10 % for HL) and the exact match value is close to 85 %. In scenario 1, RF obtains an increase of +22 % compared to the exact match of NB and +10 % compared to the macro precision of DT. Moreover, in scenario 2, the reduction of the dataset has a slight non-critical effect on the classifiers.

---

[17]Available at https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html, October 2024.

[18]Available at https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html, October 2024.

[19]Available at https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html, October 2024.

Compared to the most closely related work from the literature, even though they exploit data from low-cost activity trackers, the multi-label approach by Lee et al. [2] attains similar results. However, as explained in the related work discussion, the authors focused on mild cognitive impairment. Moreover, the multi-label strategy is also different as they applied the binary relevance method with single-label classifiers for anxiety and depression separately, not considering the correlation between these two conditions. In contrast, we leverage the MTS. Moreover, the experimental data of this work is limited to 20 samples, while our approach results in a more realistic and transparent assessment using free dialogues.

Table 4
ML classification results.

Scenario	Model	Acc.	Exact match	Precision		Recall		F-measure		HL
				Macro	Micro	Macro	Micro	Macro	Micro	
1	NB	72.71	62.95	73.40	77.77	72.49	77.42	72.91	75.97	27.29
	DT	83.46	76.03	83.48	87.03	**84.53**	87.33	83.98	85.94	16.54
	RF	**89.43**	**84.72**	**93.50**	**93.39**	84.18	**90.19**	88.03	**91.00**	**10.57**
2	NB	72.61	59.03	70.23	76.97	75.82	81.85	72.91	77.14	27.39
	DT	86.62	78.52	84.69	89.84	**86.96**	**91.49**	85.81	89.32	13.38
	RF	**88.33**	**84.21**	**92.56**	**91.70**	83.76	89.09	**87.35**	**89.70**	11.67

4.6. Explainability

Figure 2 shows the dashboard accessible to the caregivers, physicians, and end users with the four most relevant features from Table 2 on the top. The boxes are green if the feature values are below 50 % and red otherwise. At the bottom of the figure, the explanation generated by the `gpt-4o-mini` model is shown. Our approach uses the prompt engineering template described in Listing 5 filled with the values of the average and the Q_2 of the features in Table 2, the conversations of the last 2 sessions, and the predicted majority category. On the right, the dashboard indicates the prediction and the confidence percentage using `Predict_Proba` function[20].

LISTING 5: Prompt for explainability.

```
This system analyzes the user's anxiety and depression state using conversations
with a chatbot. The last 30 sessions of the user return the following feature
values:

Average:
    insecurity:X,
    loneliness:X,
    negative_emotion:X,
    positive_emotion:X,
    sadness:X,
    anguish:X,
    health_issues:X,
    catastrophic_terms:X,
    emphasized_terms:X,
    repeated_concepts:X,
    interjections:X,
    negative_adverbs:X,
    negatives_terms:X,
    polarity:X

Q2:
```

---

[20] Available at https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html, October 2024.

```
        insecurity :X,
        loneliness :X,
        negative_emotion :X,
        positive_emotion :X,
        sadness :X,
        anguish :X,
        health_issues :X,
        catastrophic_terms :X,
        emphasized_terms :X,
        repeated_concepts :X,
        interjections :X,
        negative_adverbs :X,
        negatives_terms :X,
        polarity :X

Moreover , the last 2 conversations are :

-----------------------------------------
[ conversations ]
-----------------------------------------

The prediction of our machine learning model is that the user <does not suffer |
suffers > [ anxiety / depression / anxiety and depression ].

Generate an exposition of no more than 400 characters in natural language that
summarizes the reasons why this prediction has been generated by the model based on
the information provided .
```

5. Conclusion

Given the appalling consequences of anxiety and depression, timely detection of these conditions is of uttermost importance. Traditional screening methods are time-consuming and rely on rigid subjective assessment with interviews and questionnaires. Moreover, despite the strong relationship between anxiety or stress and depression, few studies address the joint assessment of several conditions.

AI-based solutions have proposed several advantages regarding flexibility, scalability, and personalization. However, their performance in specific classification problems with task-specific data like anxiety and depression is still immature when used as final solutions, as is the case with LLMs. Moreover, some solutions lack generalization and multitask robustness, apart from low interpretability, which prevents their practical use beyond academic research. Interpretability and explainability are especially relevant in this field, provided their direct impact on clinicians' decision-making and, thus, the patient's well-being.

In this work, an entirely novel system for the multi-label classification of anxiety and depression is proposed. Another relevant contribution lies in using LLMs for feature extraction, which are intrinsically explicable but lack specific downstream knowledge, with ML models operating in a multi-label setting, which can offer higher accuracy but lack explainability. Specifically, relying on LLMs solely as part of the feature engineering module to extract user-level knowledge from free dialogues with a conversational assistant, we tackle the hallucination problem. In addition, we leverage formal medical knowledge using clinical scales for anxiety and depression to label the experimental data. Moreover, explainability descriptions of the model's decision are provided in a graphical dashboard along with the confidence of the results to promote the solution's trustworthiness, reliability, and accountability. Experimental results on a real dataset attain 90% accuracy, improving those in the prior literature. The ultimate objective is to contribute in an accessible and scalable way before formal treatment occurs in the healthcare systems.

In future work, we plan to evolve the solution to study severity levels of mental health conditions, as well as deploy the system in a real-world setting (*i.e.*, stream-based ML). Another line of work will focus on the analysis of non-verbal and paraverbal data (*e.g.*, voice modulation).

Figure 2: Explainable dashboard.

Declaration of competing interest

The authors have no competing interests to declare relevant to this article's content.

Declaration of studies in humans

This study was carried out following the World Medical Association Declaration of Helsinki.

Acknowledgments

References

[1] T. Ahmed, S. Ivan, A. Munir, S. Ahmed, Decoding depression: Analyzing social network insights for depression severity assessment with transformers and explainable AI, Natural Language Processing Journal 7 (2024) 100079. doi:`10.1016/j.nlp.2024.100079`.

[2] T. R. Lee, G. H. Kim, M. T. Choi, Identification of Geriatric Depression and Anxiety Using Activity Tracking Data and Minimal Geriatric Assessment Scales, Applied Sciences (Switzerland) 12 (2022) 2488. doi:`10.3390/app12052488`.

[3] B. Saylam, Ö. D. İncel, Multitask Learning for Mental Health: Depression, Anxiety, Stress (DAS) Using Wearables, Diagnostics 14 (2024) 501. doi:`10.3390/diagnostics14050501`.

[4] S. Marwaha, E. Palmer, T. Suppes, E. Cons, A. H. Young, R. Upthegrove, Novel and emerging treatments for major depression, The Lancet 401 (2023) 141–153. doi:`10.1016/S0140-6736(22)02080-3`.

[5] F. B. Tumaliuan, L. Grepo, E. R. Jalao, Development of Depression Data Sets and a Language Model for Depression Detection: Mixed Methods Study, JMIR Data 5 (2024) e53365. doi:10.2196/53365.

[6] X. Wang, K. Liu, C. Wang, Knowledge-enhanced Pre-Training large language model for depression diagnosis and treatment, in: Proceeding of IEEE International Conference on Cloud Computing and Intelligence Systems, IEEE, 2023, pp. 532–536. doi:10.1109/CCIS59572.2023.10263217.

[7] A. K. Chowdhury, S. R. Sujon, M. S. S. Shafi, T. Ahmmad, S. Ahmed, K. M. Hasib, F. M. Shah, Harnessing large language models over transformer models for detecting Bengali depressive social media text: A comprehensive study, Natural Language Processing Journal 7 (2024) 100075. doi:10.1016/j.nlp.2024.100075.

[8] A. Nowacki, W. Sitek, H. Rybiński, LLMental: Classification of Mental Disorders with Large Language Models, in: Proceedings of the International Symposium on Methodologies for Intelligent Systems, Springer, 2024, pp. 35–44. doi:10.1007/978-3-031-62700-2_4.

[9] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, E. Cambria, MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare, in: Proceedings of the Language Resources and Evaluation Conference, European Language Resources Association, 2022, p. 7184–7190.

[10] K. Yang, T. Zhang, Z. Kuang, Q. Xie, J. Huang, S. Ananiadou, MentaLLaMA: interpretable mental health analysis on social media with large language models, in: Proceedings of the ACM on Web Conference, Association for Computing Machinery, 2024, pp. 4489–4500. doi:10.1145/3589334.3648137.

[11] V. Vajre, M. Naylor, U. Kamath, A. Shehu, PsychBERT: A Mental Health Language Model for Social Media Mental Health Behavioral Analysis, in: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, IEEE, 2021, pp. 1077–1082. doi:10.1109/BIBM52615.2021.9669469.

[12] T. Wang, A. Roberts, D. Hesslow, T. L. Scao, H. W. Chung, I. Beltagy, J. Launay, C. Raffel, What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization?, in: Proceedings of Machine Learning Research, volume 162, MLR Press, 2022, pp. 1–21.

[13] M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, P. Rajpurkar, Foundation models for generalist medical artificial intelligence, Nature 616 (2023) 259–265. doi:10.1038/s41586-023-05881-4.

[14] Y. Liu, X. Ding, S. Peng, C. Zhang, Leveraging ChatGPT to optimize depression intervention through explainable deep learning, Frontiers in Psychiatry 15 (2024) 1383648. doi:10.3389/fpsyt.2024.1383648.

[15] X. Xu, B. Yao, Y. Dong, S. Gabriel, H. Yu, J. Hendler, M. Ghassemi, A. K. Dey, D. Wang, Mental-llm: Leveraging large language models for mental health prediction via online text data, in: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, volume 8, Association for Computing Machinery, 2024, pp. 1–32. doi:10.1145/3643540.

[16] N. Gomes, M. Pato, A. R. Lourenço, N. Datia, A Survey on Wearable Sensors for Mental Health Monitoring, Sensors 23 (2023) 1330. doi:10.3390/s23031330.

[17] A. M. Salih, Z. Raisi-Estabragh, I. B. Galazzo, P. Radeva, S. E. Petersen, K. Lekadir, G. Menegaz, A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME, Advanced Intelligent Systems (2024) 2400304. doi:10.1002/aisy.202400304.

[18] L. Coroama, A. Groza, Explainable Artificial Intelligence for Person Identification, in: Proceedings of the IEEE International Conference on Intelligent Computer Communication and Processing, IEEE, 2021, pp. 375–382. doi:10.1109/ICCP53602.2021.9733525.

[19] L. Nannini, J. Alonso-Moral, A. Catala, M. Lama, S. Barro, Operationalizing Explainable AI in the EU Regulatory Ecosystem, IEEE Intelligent Systems (2024) 37–48. doi:10.1109/MIS.2024.3383155.

[20] L. Ren, H. Lin, B. Xu, S. Zhang, L. Yang, S. Sun, Depression Detection on Reddit With an Emotion-Based Attention Network: Algorithm Development and Validation, JMIR Medical Informatics 9 (2021) e28754. doi:10.2196/28754.

[21] A. B. S. Rahman, H.-T. Ta, L. Najjar, A. Azadmanesh, A. S. Gönul, DepressionEmo: A novel dataset for multilabel classification of depression emotions, Journal of Affective Disorders (2024) 445–458. doi:10.1016/j.jad.2024.08.013.

[22] R. W. Levenson, Stress and Illness: A Role for Specific Emotions, Psychosomatic Medicine 81 (2019) 720–730. doi:10.1097/PSY.0000000000000736.

[23] S. A. Qureshi, G. Dias, M. Hasanuzzaman, S. Saha, Improving Depression Level Estimation by Concurrently Learning Emotion Intensity, IEEE Computational Intelligence Magazine 15 (2020) 47–59. doi:10.1109/MCI.2020.2998234.

[24] S. Ghosh, A. Ekbal, P. Bhattacharyya, What Does Your Bio Say? Inferring Twitter Users' Depression Status From Multimodal Profile Information Using Deep Learning, IEEE Transactions on Computational Social Systems 9 (2022) 1484–1494. doi:10.1109/TCSS.2021.3116242.

[25] E. Turcan, S. Muresan, K. McKeown, Emotion-Infused Models for Explainable Psychological Stress Detection, in: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2021, p. 2895–2909. doi:10.18653/v1/2021.naacl-main.230.

[26] S. Ghosh, A. Ekbal, P. Bhattacharyya, A Multitask Framework to Detect Depression, Sentiment and Multi-label Emotion from Suicide Notes, Cognitive Computation 14 (2022) 110–129. doi:10.1007/s12559-021-09828-7.

[27] S. Sarkar, A. Alhamadani, L. Alkulaib, C.-T. Lu, Predicting Depression and Anxiety on Reddit: a Multi-task Learning Approach, in: Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, IEEE, 2022, pp. 427–435. doi:10.1109/ASONAM55673.2022.10068655.

[28] D. Park, G. Lee, S. Kim, T. Seo, H. Oh, S. J. Kim, Probability-based multi-label classification considering correlation between labels–focusing on DSM-5 depressive disorder diagnostic criteria, IEEE Access (2024) 70289–70296. doi:10.1109/ACCESS.2024.3401704.

[29] L. Ilias, D. Askounis, Multitask learning for recognizing stress and depression in social media, Online Social Networks and Media 37-38 (2023) 100270. doi:10.1016/j.osnem.2023.100270.

[30] D. Park, S. Lim, Y. Choi, H. Oh, Depression Emotion Multi-Label Classification Using Everytime Platform With DSM-5 Diagnostic Criteria, IEEE Access 11 (2023) 89093–89106. doi:10.1109/ACCESS.2023.3305477.

[31] V. B. de Souza, J. C. Nobre, K. Becker, DAC Stacking: A Deep Learning Ensemble to Classify Anxiety, Depression, and Their Comorbidity From Reddit Texts, IEEE Journal of Biomedical and Health Informatics 26 (2022) 3303–3311. doi:10.1109/JBHI.2022.3151589.

[32] L. Ilias, S. Mouzakitis, D. Askounis, Calibration of Transformer-Based Models for Identifying Stress and Depression in Social Media, IEEE Transactions on Computational Social Systems 11 (2024) 1979–1990. doi:10.1109/TCSS.2023.3283009.

[33] J. Ohse, B. Hadžić, P. Mohammed, N. Peperkorn, M. Danner, A. Yorita, N. Kubota, M. Rätsch, Y. Shiban, Zero-Shot Strike: Testing the generalisation capabilities of out-of-the-box LLM models for depression detection, Computer Speech & Language 88 (2024) 101663. doi:10.1016/j.csl.2024.101663.

[34] Y. Wang, D. Inkpen, P. K. Gamaarachchige, Explainable depression detection using large language models on social media data, in: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology, Association for Computational Linguistics, 2024, pp. 108–126.

[35] A. Rivolli, J. Read, C. Soares, B. Pfahringer, A. C. P. L. F. de Carvalho, An empirical analysis of binary transformation strategies and base algorithms for multi-label learning, Machine Learning 109 (2020) 1509–1563. doi:10.1007/s10994-020-05879-3.