# TextualTherapists at MentalRiskES-IberLEF2023: Early Detection of Depression using a User-level Feature-based Machine Learning Approach

Alberto Fernández-Hernández[1], Raúl Moreno-Sánchez[2], José Viosca-Ros[3], Raquel Enrique-Guillén[4], Noa Patricia Cruz-Díaz[5] and Salud María Jiménez-Zafra[6,*]

[1]*Enterprise Business Unit Department, Vodafone Group Plc, 28042, Spain*
[2]*OpenSpring IT IBERIA S.L., 28036, Spain*
[3]*SoGooData (data for social good), 28025, Spain*
[4]*CaixaBank Business Intelligence, 28050, Spain*
[5]*Intrum Global Technologies Spain, 28033, Spain*
[6]*Computer Science Department, SINAI, CEATIC, Universidad de Jaén, 23071, Spain*

Abstract

This paper presents the participation of the TextualTherapists team in the MentalRiskES shared task at the IberLEF 2023 evaluation campaign. This shared task focuses on the early risk prediction of mental disorders in Spanish. Specifically, we have participated in Task 2.a on detecting whether a user suffers from depression or not based on a set of comments posted on Telegram. We addressed this task using a machine learning approach that integrates lexical, sentiment, toxicity, and emotional features and that takes into account the PHQ9 Patient Questionnaire. There was a total of 33 runs submitted by the participating teams. The best run sent by our team placed in position 4th for depression detection, with a Macro-F1 score of 0.729, and position 6th for early risk depression detection with an ERDE-30 of 0.159, being 0.737 and 0.140 the best result obtained in the competition, respectively.

**Keywords**

Depression detection, early risk prediction of depression, machine learning approach, PHQ9 Patient Questionnaire, lexical features, sentiment features, toxicity features, emotional features

## 1. Introduction

Depression is a prevalent mental health disorder that affects millions of individuals worldwide. According to the World Health Organization, 3.8% of the population, approximately 280 million people in the world, experience depression. In addition, depression is a leading cause of disability around the world and contributes greatly to the global burden of disease [1]. Consequently, its

*IberLEF 2023, September 2023, Jaén, Spain*
*Corresponding author.
albertofernandezh98@gmail.com (A. Fernández-Hernández); raul.moreno@openspring.com (R. Moreno-Sánchez); jviosca@gmail.com (J. Viosca-Ros); raquelen12@hotmail.com (R. Enrique-Guillén); noa.cruz@intrum.com (N. P. Cruz-Díaz); sjzafra@ujaen.es (S. M. Jiménez-Zafra)
0000-0002-5645-6475 (J. Viosca-Ros); 0009-0007-7159-0908 (R. Enrique-Guillén); 0000-0002-6685-6747 (N. P. Cruz-Díaz); 0000-0003-3274-8825 (S. M. Jiménez-Zafra)

impact on individuals' well-being, quality of life, and overall societal productivity necessitates the development of efficient and accurate methods for early detection.

Underdiagnosis of depression is becoming of growing concern worldwide. However, automated user data analysis can provide a helpful pre-screening step to boost early detection of depressive signs. Indeed, recent advancements in Natural Language Processing (NLP) techniques have shown its potential to help detect signs of depression from textual data. In fact, recently organized workshops and shared tasks have fostered discussions around detection of depression from social media texts such as the Early-Risk Identification task (eRisk) hosted at Cross-Lingual Evaluation Forum (CLEF) during the last few years [2, 3, 4] or the DepSign-LT-EDI@ACL-2022 shared task [5].

Unfortunately, these campaigns have focused mainly on English, leaving aside other languages like Spanish. MentalRiskES [6], organized at the Iberian Languages Evaluation Forum (IberLEF 2023) [7], attempts to mitigate this gap by proposing a novel task to early detect the risk of mental disorders in Spanish comments from Telegram users, including eating disorders and depression. This paper describes our participation in Task 2.a of this competition whose aim is to detect whether users suffer from depression.

The remainder of the paper is structured as follows. Section 2 presents the task, the dataset and the evaluation measures. Section 3 gives a detailed explanation of the methodology used to develop our proposal for detecting depression in texts. Section 4 describes the corresponding experimental settings. The results are discussed in Section 5, along with the error analysis. Finally, Section 6 draws some conclusion and future work.

## 2. Task description

The MentalRiskES shared task [6] comprises Task 1: eating disorders detection, Task 2: depression detection and Task 3: unknown disorder detection. We focused on Task 2.a which consists of detecting whether users suffer from depression (binary classification). A user is considered to be suffering from depression when expresses everyday situations, desires, or actions related to the suffering of such pathology.

The task must be resolved as an online problem, that is, the proposed systems must be able to detect a potential risk as early as possible in a continuous stream of data. Therefore, the performance not only depends on the accuracy of the systems but also on how fast the problem is detected.

### 2.1. Dataset

The dataset used in Task 2.a consists of Spanish messages from conversations of 334 users in public Telegram groups. These messages were anonymized and manually annotated through the Prolific service, where each user's history were labelled by ten annotators with the label 0 for "control" (negative, the user does not suffer from depression) or 1 for "suffer" (positive). The probability of a disorder were established by dividing the number of annotators that found evidence of suffering from the targeted disorder by the number of total annotators. A value of 0 means 100% negative and a value of 1 would be 100% positive. The distribution of the dataset for this task is presented in Table 1.

**Table 1**
Task 2.a dataset statistics

| | trial | train | test | total |
|---|---|---|---|---|
| **suffer** | 6 | 94 | - | 100 |
| **control** | 4 | 81 | - | 89 |
| **total** | 10 | 175 | 149 | 334 |

### 2.2. Evaluation

The proposed systems are evaluated using different types of performances depending on the task. We include those of Task 2.a, which is the one in which we have participated:

- Task 2.a.
  - Binary classification: accuracy, micro, and macro precision, recall, f-score
  - Latency-based: early risk detection metric ERDE or its variants

Efficiency metrics are also used in order to measure the impact of the system in terms of resources needed and environmental issues. They include the following information:

- Total RAM needed
- Total % of CPU usage
- Floating Point Operations per Second (FLOPS)
- Total time to process (in milliseconds)
- Kg in CO2 emissions. For this, the Code Carbon tool will be used.

## 3. Methodology

In this section, we present the methodology employed in our study for detecting depression in texts using machine learning models. By delineating the methodology, we aim to provide a comprehensive understanding of the techniques and procedures used in our research, ensuring transparency and reproducibility. First, we outline the Exploratory Data Analysis (EDA) conducted and, later, we describe our system proposal.

### 3.1. Exploratory Data Analysis

We start the EDA by examining the information of the dataset. Trial and train sets consist of 4 columns: i) id_message: message unique identifier, ii) message: text message, iii) date: publication date and, iv) user_id: user unique identifier. In addition, a gold file is provided for each set in which each user_id is associated with a label: 0 for "control" and 1 for "suffer". The number of users and the number of messages supplied in each set can be seen in Table 1 and Table 2, respectively. It should be noted that messages are not annotated, only users.

Analyzing the number of messages per user (Figure 1) and the text length (Figure 2) on train data, it can be seen that:

- 80 % of users have 40 messages or less, approximately.
- 80 % of users have a text length of  645 or less.

**Table 2**
Total messages per set

| set | total messages |
|---|---|
| trial | 624 |
| train | 6248 |
| **total** | 6872 |

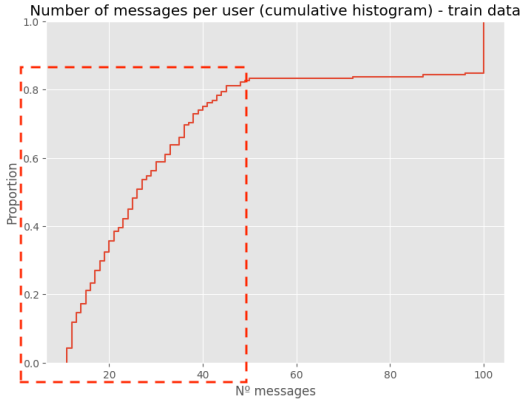


**Figure 1:** Number of messages per user on train data (cumulative histogram)

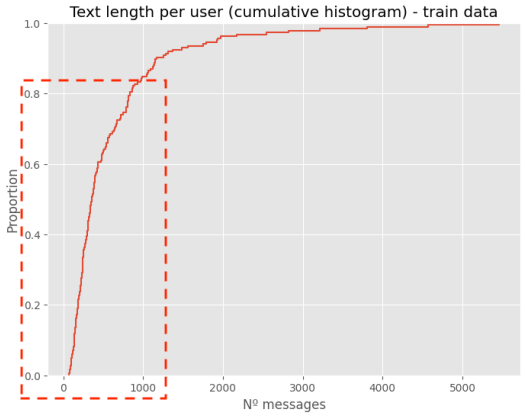


**Figure 2:** Text length per user on train data (cumulative histogram)

From the analysis, it can be concluded that there is a limited number of messages per user. This fact coupled with the absence of message-level labels, leads us to believe that a user-level

approach complemented with message-level feature engineering could be a viable solution for detecting depression in text messages. By grouping all messages by user_id as input for our model, we can leverage the available data more effectively, capturing general behavioral patterns and linguistic cues to identify signs of depression. It is important to note that, in this dataset, we cannot infer which messages specifically mention depression or serve as strong indicators of depression, as the dataset is labeled at the user_id level.

Moreover, during the EDA phase, we also noticed that emojis are in text format. For instance: ☹ as "sad face" or "cara triste" in Spanish. Emojis have become an increasingly prevalent form of communication in digital platforms, offering an additional layer of emotional expression to textual content. By leveraging the expressive power of emojis, we aim to enhance the effectiveness of our model in capturing emotional features associated with depression. To do so, we utilize a text-to-emoji conversion database that maps textual content to their corresponding emoji representations (https://emojiterra.com/es/).

## 3.2. System proposal

In this section, we present our system proposal for detecting depression in texts. After discussing in the previous section how data should be organized, we now focus on the feature engineering phase. Here, we explore the variables that were carefully selected and incorporated into our machine-learning model, aiming to create a robust and accurate system.

### 3.2.1. Text translation

Before performing the feature engineering we translated the Spanish dataset into English since English is the dominant language in the NLP field, making it possible to access a larger amount of resources: tools, libraries, transformers, etc. The chosen solution is the googletrans library, which employs the Google Translate Ajax API to execute functions such as language detection and translation: https://github.com/ssut/py-googletrans.

### 3.2.2. Feature engineering

The process of feature engineering plays a critical role in developing effective machine learning models. By identifying and extracting meaningful patterns and characteristics from the textual data, we can provide the model with valuable insights. In this subsection, we aim to transform the raw text inputs into a set of chosen features that capture the essence of depressive messages.

To do this, we carefully select a group of variables, based on extensive research in psychology, psychiatry, and NLP [8, 9, 10, 11]. Our objective was to identify the most relevant textual clues and linguistic indicators that could serve as predictive features for depression detection.

#### Empath features

In our research, we have taken into consideration various studies that utilize lexical features [12, 13, 14, 15]. When it comes to lexical analysis, the well-known option is the Linguistic Inquiry and Word Count (LIWC) software, which has been widely used in various fields: from psychology, linguistics, and communications to social sciences. However, LIWC is proprietary software that requires purchasing a license. Consequently, we have decided to incorporate

an open-source approach, named *empath* library. This library serves as a valuable tool for analyzing text across different lexical categories, by counting the occurrences for each "empath feature" term (https://github.com/Ejhfast/empath-client).

Chosen features include a wide range of topics: alcohol, hate, envy, health, nervousness, weakness, horror, suffering, kill, fear, friends, sexual, body, family, irritability, violence, sadness, disgust, exasperation, emotional, anger, poor, pain, timidity, cheerfulness,medical_emergency, rage,positive_emotion, negative_emotion, ugliness, weapon, shame, torment, help, office, sleep, money, school, home, hygiene, phone, work, appearance, optimism, youth, joy, white_collar_job, morning, night, college, sports, neglect, disappointment, children, contentment, music, musical, deception, blue_collar_job, clothing, valuable, swearing_terms, and exercise.

**Part of Speech features**

Part Of Speech tagging (POS) is performed by counting the occurrences for each "empath feature" term; including adjectives, superlatives, adverbs, verbs, nouns, and past tense verbs. Additional studies reveal patterns between POS tags and depression: individuals diagnosed with depression tend to use fewer common and proper nouns in comparison with control users, along with more verbs and adverbs in their posts [8]; suggesting that the use of linguistic styles such as pronouns and articles provide information about how individuals respond to psychological triggers [9, 10].

**PHQ-9 terms**

To incorporate clinical tools commonly used by mental health professionals to diagnose depression, we quantified the occurrence of written expressions indicative of depression signs, mimicking what is done by mental health professionals when they use the PHQ-9, which is the depression module of the Patient Health Questionnaire (PHQ). The PHQ-9 [16] is an easy to use and self-administered patient questionnaire that scores the frequency of occurrence (from absent to nearly every day) of each of the nine diagnostic criteria for depression described in the Diagnostic and Statistical Manual of Mental Disorders (DSM), the most widely accepted nomenclature and reference book used by clinicians and researchers for the classification of mental disorders [17]. These diagnostic criteria are the following: anhedonia, concentration issues, eating issues, fatigue, mood problems, psychomotor problems, self-esteem issues, self-harm and sleep problems (we also included panic attacks as a 10th criterion, as they frequently co-occur with depressive disorders in the same patients [17]). We looked for signs suggestive of each of these criteria by using regular expressions, some of which were extracted from a GitHub repository (https://github.com/thongnt99/acl22-depression-phq9) containing expressions based on the work of [11].

**Sentiment features**

We compute different sentiment features from a variety of well-known libraries, models and tools: TextBlob [18], VADER [19], RoBERTa [20] and the Emoji Sentiment Ranking [21]. TextBlob was selected in order to extract the polarity and subjectivity of the messages and VADER to get sentiment scores. Differently from TextBlob, VADER focuses on elements that typically appear in social media such as emojis, repetitive words, and punctuations (exclamation marks,

for example). Despite the massive use of VADER library, it has some flaws: i) it is sensible to grammatical structure and ii) it uses pre-defined dictionaries with pre-defined scores. Thus, we decided to add other sentiment scores, based on transformers, specifically from RoBERTa base sentiment model. Finally, we also incorporate sentiment scores values for emojis using the Emoji Sentiment Ranking.

1. TextBlob polarity
2. TextBlob subjetivity
3. VADER positive sentiment score
4. VADER neutral sentiment score
5. VADER negative sentiment score
6. RoBERTa positive sentiment score
7. RoBERTa neutral sentiment score
8. RoBERTa negative sentiment score
9. Number of "positive" emojis
10. Number of "neutral" emojis
11. Number of "negative" emojis
12. Average sentiment score (based on previous emojis features)

**Toxicity features**
We also take into account the following toxicity features: toxic, severe_toxic, obscene, threat, insult and identity_hate, extracted from the transformer model "toxic-bert" [22], available on HuggingFace: https://huggingface.co/unitary/toxic-bert.

**Emotional features**
To detect emotion in texts, we use NRClex [23], a library which predicts the sentiments and emotion of a given text. The package contains approximately 27,000 words and is based on the National Research Council Canada (NRC) affect lexicon and the NLTK library's WordNet synonym sets [24]. We include the emotion value expressed in texts for fear, anger, anticipation, trust, surprise, sadness, disgust, joy and, positive and negative sentiment values.

**Readability features**
Readability refers to the ease with which a reader can understand a given text. It is influenced by factors such as sentence structure, word complexity, and overall linguistic coherence. By assessing the readability of text, the objective is to determine whether text complexity and readability can be used to distinguish between depressed and non-depressed texts. To do so, several readability features are included:

1. Kincaid Index [25]: This index is extensively used in the field of education. It is a readability metric that quantifies the difficulty level of a text based on its average sentence length and average number of syllables per word. It measures the grade level required to comprehend the text.

2. ARI Index [26]: The automated readability index (ARI) measures the comprehension level required to understand a text based on factors such as sentence length and word complexity.
3. Coleman Index [27]: It is an alternative to other readability formulas which computes the U.S. grade level required to comprehend a text. One advantage of the Coleman Index is that it does not rely on syllable count, making it simpler to calculate compared to some other readability metrics. However, it does assume that longer sentences and words with more characters are indicative of more difficult texts.
4. Gunning-Fog Index [28]: It provides an estimate of the years of formal education required to understand a text easily. The index is commonly used by writers, editors, and educators to evaluate the complexity of written materials.
5. LIX [29]: It calculates the difficulty of reading a foreign text.
6. SMOG [30]: The Simple Measure of Gobbledygook is designed to estimate the years of education an individual needs to easily understand a piece of writing. The SMOG grade formula involves counting the number of polysyllabic words in a sample text. The formula assumes that the more polysyllabic words there are in a text, the more difficult it is to comprehend.
7. Dale Chall [31]: It is a readability assessment tool used to measure the readability of a written text. The formula uses two main components: the average sentence length and the percentage of words that are not on the Dale-Chall word list, knowns as “difficult words”. The Dale-Chall word list consists of around 3,000 familiar words that are commonly used in everyday English. It excludes technical or domain-specific vocabulary that may be unfamiliar to most readers.

**Other features**

Finally, we also explored other general features. These additional variables, such as text length, number of complex words, and other linguistic elements, could play an important role in distinguishing between individuals who may or may not be diagnosed with depression:

1. Number of first person pronouns singular (”I”, ”me” and ”mine”)
2. Number of words
3. Number of words in uppercase
4. Number of sentences
5. Number of paragraphs
6. Number of long words: greater or equal than 7 characters
7. Number of complex words
8. Number of interrogative (¿?) and exclamation signs (¡!)
9. Number of emoticons (e.g :) ;) :D D:)
10. Number of quantifiers: **some, several, a number of, enough, numerous, plenty of, a lot of, lots of, much, many, few, little**

# 4. Experimental setup

Once the above features have been extracted from the dataset, the next step is to organize and group those features. An important aspect to keep in mind is that the dataset used for this task is labeled at the user level and not at the individual message level: the labeling is done based on the overall mental health status of the user.

When grouping features by the user_id column, it is essential to take into account the nature of the variables being analyzed. Different groups of features require different aggregation methods to ensure meaningful and logical representations of these variables. In our approach, we use two main aggregation methods: summation and averaging.

1. Sum Aggregation: features that represent counts or frequencies or lexical elements are aggregated using the sum method. For instance, Part of Speech (PoS) features, such as the number of nouns or verbs in each sentence. By summing them, we obtain an overall count of nouns or verbs used by a user across all their messages. We employ this strategy for empath, Part of Speech (PoS), PHQ-9 terms and other features.
2. Mean Aggregation: on the other hand, features based on probability outputs or continuous scales are aggregated using the mean method. For instance, probability values typically range from 0 to 1. Consequently, summing probabilities can lead to values that exceed the intended range and distort the interpretation of the feature. We use this method for readability, sentiment, emotion and toxicity features.

## 4.1. Modelling using training data

The training process of the model involved several steps and techniques to ensure optimal performance and accuracy. In this section, we will delve into the methodology used to train the model, including the libraries and techniques employed.

### 4.1.1. AutoML library

To streamline the training process and leverage its powerful machine learning capabilities, we utilized the PyCaret library [32]: an open-source, low-code machine learning library in Python that automates various steps in the machine learning workflow, including data preprocessing, feature selection, model training, hyperparameter tuning, and evaluation.

### 4.1.2. K-Fold Cross-Validation

Since we had a limited amount of data available for training, we employed K-Fold cross-validation to obtain reliable performance estimates. K-Fold cross-validation is a technique that partitions the dataset into K subsets or folds of approximately equal size. The model is trained and evaluated K times, each time using a different fold as the validation set and the remaining folds as the training set. This approach helps mitigate the risk of overfitting and provides a more robust evaluation of the model's performance. For this task, we have used the default number of folds (K): 10.

#### 4.1.3. Feature Selection

To enhance the model's predictive capabilities and reduce the dimensionality of the dataset, we employed feature selection techniques. In particular, we utilized the SelectFromModel function from the scikit-learn library [33]: http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectFromModel.html. This method allows us to select the most important features by training an estimator and extracting the top-ranked features based on their importance scores. By doing so, we can focus on the most relevant features, which often leads to improved model performance.

#### 4.1.4. Model Selection

After performing feature selection, we trained multiple machine learning models using PyCaret's automated workflow. PyCaret supports a wide range of algorithms and provides a convenient way to compare their performance on the dataset: Gradient Boosting Classifier, Dummy Classifier, CatBoost Classifier, Random Forest Classifier, Decision Tree Classifier, Light Gradient Boosting Machine, Logistic Regression, Extreme Gradient Boosting, Ada Boost Classifier, Quadratic Discriminant Analysis and K Neighbors Classifier.

To select the top-performing models, we sorted them based on their F1 scores, a common metric used in classification tasks that balances precision and recall. The top three models were chosen for further evaluation and comparison.

Overall, the training process involved using PyCaret's automated workflow to preprocess the data, perform K-Fold cross-validation, split the dataset, apply feature selection, and train multiple models. The top-performing models were then selected based on their F1 scores, setting the stage for subsequent evaluation and fine-tuning of the models to achieve optimal results.

### 4.2. Test data inference: early detection

Once top models are trained, as test data is released in rounds, inference is applied on test data in a particular way:

1. For each round, data is retrieved in JSON file, with same format as train data.
2. Once it is retrieved, translation and feature engineering are applied for each message.
3. Next, features are grouped by user id, following same format as training data.
4. Then, we load top three models to predict labels, as well as efficiency metrics, including:
    - RAM needed
    - Percentage of CPU usage
    - Floating Point Operations per Second (FLOPS)
    - Total time to process (in milliseconds)
    - Kg in CO2 emissions. For this, the Code Carbon tool is used: https://pypi.org/project/codecarbon/
5. Finally, we submit predictions for each model.
   We iterate over this loop until maximum round is reached (i.e, there is no data to retrieve). Thus, it acts as an "early detection" of depression iterator, checking in which round a user has been diagnosed with depression.

## 5. Results and discussion

After completing the AutoML training, we obtained the three best results on our training dataset, sorted by F1.

### 5.1. Results on KFold train

In Table 3, we present the results of the top-3 models for depression detection task using 10-fold cross-validation in the traning data. Top models include Random Forest Classifier, Light Gradient Boosting Machine, and Logistic Regression:

**Table 3**
Results on the training phase using 10-fold cross-validation

| model | accuracy | AUC | recall | precision | F1 | training time (sec.) |
|---|---|---|---|---|---|---|
| Random Forest Classifier | **0.8181** | **0.8915** | **0.8250** | **0.8461** | **0.8306** | 0.2580 |
| Light Gradient Boosting Machine | 0.7771 | 0.8792 | **0.8250** | 0.8014 | 0.8004 | 0.1280 |
| Logistic Regression | 0.7710 | 0.8190 | 0.8125 | 0.7842 | 0.7924 | **0.1160** |

Overall, the Random Forest Classifier was the top-performing model, exhibiting high accuracy, AUC, recall, precision, and F1 score. However, the Light Gradient Boosting Machine and Logistic Regression models also show similar performance, albeit with slightly lower metrics. Table 4, Table 5 and Table 6 show the hyper parameter settings for these three models.

**Table 4**
Hyper parameters: Random Forest Classifier

| parameter | value |
|---|---|
| bootstrap | True |
| ccp_alpha | 0.0 |
| criterion | gini |
| max_depth | None |
| max_leaf_nodes | None |
| max_samples | None |
| min_impurity_decrease | 0.0 |
| min_samples_leaf | 1 |
| min_samples_split | 2 |
| min_weight_fraction_leaf | 0.0 |
| n_estimators | 100 |
| oob_score | False |

### 5.2. Feature importance

Now, we analyze feature importance of the best model (Random Forest). As it can be seen in Figure 3, there are several variables that play a moderate role, such as “Empath” features like

‘pain’, ‘nervousness’, ‘sadness’, ‘violence’, ‘suffering’, ’money’, ’youth’ and Part of Speech features: ’fe_pos_advs’ or ’Number of adverbs’. While these features contribute to the predictions, their impact is not as strong as others. However, there are features that significantly affect the predictions. Notably, “transformers-based” features have a substantial impact on the final predictions, specifically:

- Average negative sentiment score per user (‘fe_roberta_base_sentiment_negative_mean’) and average neutral sentiment score (‘fe_roberta_base_sentiment_neutral_mean’).
- Emotion features, including average optimism score (‘fe_distilbert_emotion_optimism_mean’) and average sadness score (‘fe_distilbert_emotion_sadness_mean’ and ‘fe_nrclex_emotion_sadness_mean’).
- Toxicity features: Average “insulting” score or ‘fe_insult_mean’.

Understanding the importance of these features allows us to comprehend that features obtained through transformer models’ outputs far outweigh the simplistic nature of counting variables, such as the number of Part-of-Speech (PoS) features or PHQ-9 terms. In fact, extracting features based on capturing deep contextual understanding and emotional nuances within texts (emotion, toxicity and sentiment) surpasses the limited scope of counting variables.

### 5.3. Results on test set

Once top models are trained, we evaluate their performance through test set. As shown in Table 7, Random Forest Classifier achieves the best results, with around 42.1 % correct

**Table 5**
Hyper parameters: Light Gradient Boosting Classifier

| parameter | value |
|---|---|
| boosting_type | gbdt |
| colsample_bytree | 1.0 |
| learning_rate | 0.1 |
| max_depth | None |
| min_child_samples | 20 |
| min_child_weight | 0.001 |
| min_split_gain | 0.0 |
| n_estimators | 100 |
| num_leaves | 31 |
| reg_alpha | 0.0 |
| reg_lambda | 0.0 |

**Table 6**
Hyper parameters: Logistic Regression Classifier

| parameter | value |
|---|---|
| penalty | l2 |

predictions for the top 5 individuals at highest risk of depression (ERDE5) and 16.1 % for the top 30 individuals (ERDE30). However, it took a bit longer to process, with a latency of 7,000 units per prediction (latencyTP). Nonetheless, it maintained a respectable speed score of 0.903. Overall, its performance, as measured by the latency-weighted metric, yielded a score of 0.682. Nevertheless, LightGBM model shows comparable accuracy, achieving approximately 34.2 % correct predictions for the top 5 important outcomes (ERDE5) and 16.8 % for the top 30 outcomes (ERDE30).

On the other hand, results in Table 8 show that Random Forest Classifier performed slightly lower than the best model from the competition but still achieved decent results, with an accuracy of 0.732. Moreover, this model had a macro precision of 0.766, suggesting a relatively low rate of false positives. The macro recall was 0.746, indicating that it captured a good proportion of the positive instances. The macro F1 score was 0.732, reflecting a reasonable balance between precision and recall. In any case, the difference between any metric between Random Forest and the best model is just thousandths.

Moreover, if we compare our best model with the best of the competition (in terms of

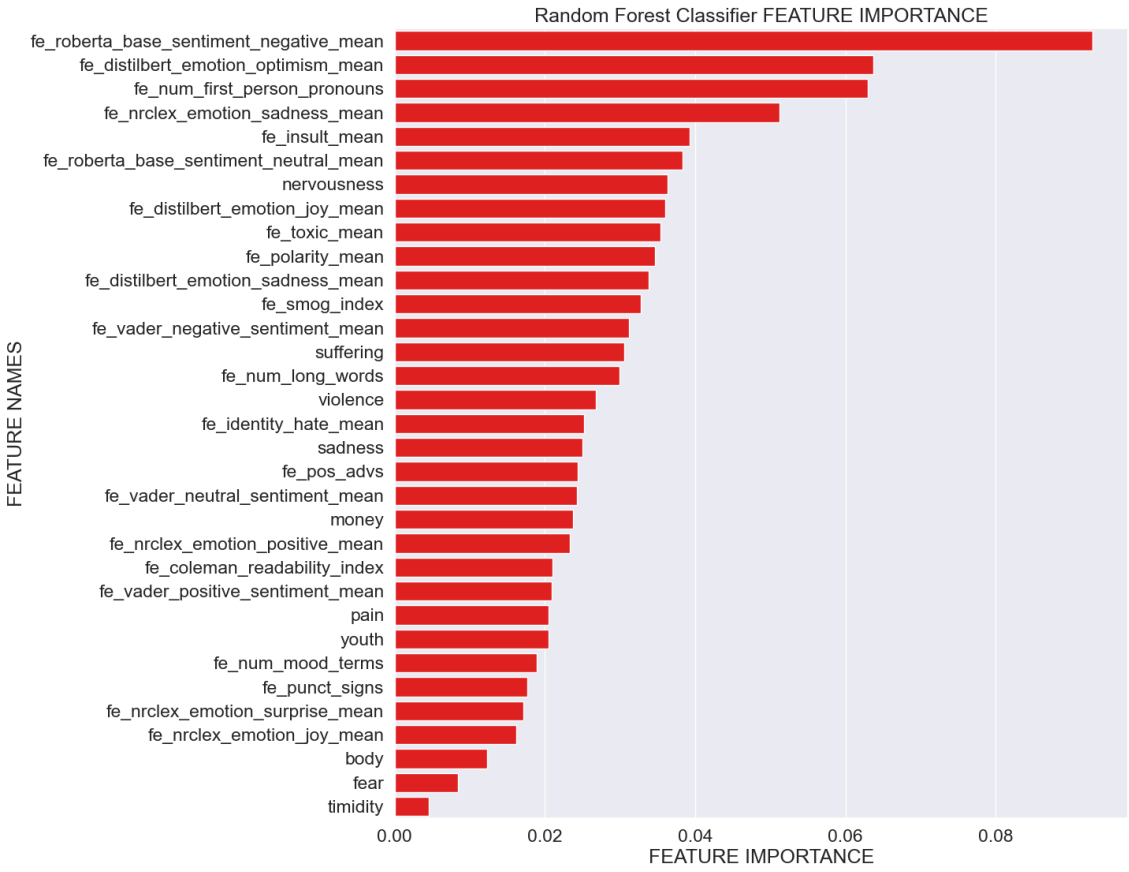


**Figure 3:** Feature importance with Random Forest

**Table 7**
Results on the evaluation phase (I)

| rank | model | ERDE5 | ERDE30 | latencyTP | speed | latency weighted |
|---|---|---|---|---|---|---|
| 6 | Random Forest Classifier | **0.421** | 0.161 | 7.000 | 0.903 | 0.682 |
| 7 | Light Gradient Boosting Machine | 0.342 | 0.168 | 3.000 | 0.967 | 0.696 |
| 17 | Logistic Regression | 0.330 | **0.205** | 2.000 | 0.984 | 0.663 |

**Table 8**
Results on the evaluation phase compared to the best model

| rank | model | accuracy | macro-precision | macro-recall | macro-F1 |
|---|---|---|---|---|---|
| 1 | Best model from the competition | **0.738** | **0.756** | **0.749** | **0.737** |
| 6 | Random Forest Classifier | 0.732 | 0.766 | 0.746 | 0.732 |
| 7 | Light Gradient Boosting Machine | 0.664 | 0.740 | 0.687 | 0.651 |
| 17 | Logistic Regression | 0.664 | 0.740 | 0.687 | 0.651 |

carbon emissions and energy consumption), we can appreciate that the best model significantly outperforms Random Forest Classifier in terms of environmental impact:

**Table 9**
Average carbon emissions and energy consumption, compared to the best model

| rank | model | Carbon emissions | CPU energy | RAM energy | Energy consumed |
|---|---|---|---|---|---|
| 1 | Best model from the competition | **5.52E-08** | **1.02E-07** | **7.73E-10** | **2.91E-07** |
| 6 | Random Forest Classifier | 3.23E-06 | 1.67E-05 | 3.15E-07 | 1.70E-05 |

The CO2 emissions for our model are more than 50 times greater than that of the best model, similar to hardware consumption. However, while most systems have been trained using both CPUs and GPUs, our best model uses only 4 CPUs with an average of CPU energy consumption of 3.56E-06 and with an average RAM energy needed of 3.15E-07. This means that our system requires fewer resources than most systems in the top 10 and is therefore more efficient in this respect. This is aligned with the CO2 emissions, with an average of 3.23E-06, making our system the third lowest environmental impact of the top 10.

In summary, both the Random Forest Classifier and Light Gradient Boosting Machine models demonstrated comparable accuracy, with the former having a slight edge, though higher energy consumption and carbon emissions.

## 5.4. Error analysis

Finally, we conducted an error analysis of the predictions with our best system (Random Forest Classifier). For this, we used the test set with the gold labels that was released by the competition organizers when publishing the competition results. First, we obtained the confusion matrix,

**Table 10**
Confusion matrix of the evaluation set

| | | Real labels | | |
|---|---|---|---|---|
| | | **suffer** | **control** | **total** |
| **Predicted labels** | **suffer** | 62 | 38 | 100 |
| | **control** | 6 | 43 | 49 |
| | **total** | 68 | 81 | 149 |

which is shown in Table 10. In it, it can be seen that of the 149 subjects to be classified, our system correctly classified 105 subjects and was wrong in predicting 44 subjects. Of the misclassified cases, 38 were false positives (FP), i.e., the system identified those subjects as suffering from depression but they had the label control. On the other hand, 6 were false negatives (FN), i.e., the system predicted those subjects as control but their real label were suffer.

We analyzed the messages of the subjects corresponding to the 6 FN and from the 38 FP, we selected the 6 predicted subjects with a higher probability. Analyzing the 6 subjects that our system predicted as control but that the dataset annotators did label as being suffering from depression (FN), we detected the following cases:

- Explicit mention of diagnosed depression, but use of positive language. In this case, we think that the system has not been able to detect this case of depression because of the use of positive language. Perhaps it would be interesting to capture explicit mentions of diagnosed depression and the weight of positive sentiments should be checked.
- Suicidal thinking. This case reflects that there are expressions denoting melancholy that our system is not capturing, probably because we make a translation of the text into English.
- Possible annotation error. The user shows signs of melancholy but without any other indicator, he/she cannot be indisputably labelled as depressed (i.e. it might be a true negative incorrectly labelled as depressed).
- Little text, but with triggering signs such as questions about treatment with antidepressants and mention of mental health professionals (psychologists). As in the first case, it would be interesting to include triggers in the system.
- Mention of having depression but conversation in positive terms, with laughter.
- Possible annotation error. Individuals offering help to another person.

As mentioned above, most of the system errors are due to an erroneous prediction of users suffering from depression (FP). In the analysis of the 6 selected subjects, we identified the following cases:

- Talking about another depressed person. Thus, it would be interesting to analyze third person verbal usage.
- Use of terms indicative of low self-esteem, but it is not clear that the user has depression.
- User who has suffered from depression in the past, but has recovered.
- User who has suffered from depression and offers help.

- User who offers help.
- User who externalizes his/her affective insecurities and his/her self-perception of gender, but has playful activity. In the system it would be interesting to give more weight to indications of playful activities because this indicates normality.

This analysis reveals important findings that could be taken into account to improve our system. It would be interesting to differentiate when users speak in the first person and when they speak in the third person, since sometimes they are talking about other people's experience. The verb tense should also be taken into account, as there are users who have suffered from depression in the past but no longer have it. On the other hand, the weight of positive sentiments should be reviewed, since in some cases it has been observed that they have had more influence than aspects of depression itself. A treatment of triggering expressions should also be included. Finally, it would be interesting to create a version of the system for Spanish texts, since it has been observed that the English system is not capable of capturing melancholic expressions in Spanish.

## 6. Conclusions and future work

In this work, we have described the details of the participation of the TextualTherapist team in Task 2.a of MentalRiskES shared task, on detecting whether users suffer from depression or not based on a set of comments they posted. We have addressed the task combining machine learning algorithms with lexical, sentiment, toxicity and emotional features, and taking into account the PHQ9 Patient Questionnaire. In conclusion, our work shows that machine learning can be used to effectively detect depression signs with relatively high efficiency. Our best model (Random Forest Classifier) achieved a macro-F1 score of 0.737. From the analyzed features we have detected that the ones that have the greatest impact on this task are those provided by the transformer models used to obtain the average negative and neutral sentiments, the average optimism and sadness score and the average insulting score.

The error analysis conducted has given us insights on how to improve our system: taking into account the person and verb tense used, incorporating a treatment of depression-triggering expressions and capturing melancholic expressions. Moreover, some possible lines for future work are the following:

- Translation and adaptation of code into Spanish (hereby bypassing the need to translate the analyzed text into English).
- Complete error analysis and detail/characterize model limitations for both early detection and detection of depression.
- Expand keywords used beyond PHQ-9 questionnaire (for example: include keywords specific to the Depression chapter of the DSM-5, include keywords indicative of known risk factors for depression such as early traumatic experiences, etc.).
- Describe the contribution of the different feature variables used in the model to the prediction of depression status (i.e. differences between suffer and control).
- Perform ablation analysis to help interpret the contribution of the different feature variables used in the model.

## Acknowledgments

This work has been partially supported by Project CONSENSO (PID2021-122263OB-C21), Project MODERATES (TED2021-130145B-I00) and Project SocialTox (PDC2022-133146-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, Project PRECOM (SUBV-00016) funded by the Ministry of Consumer Affairs of the Spanish Government, Project FedDAP (PID2020-116118GA-I00) supported by MICINN/AEI/10.13039/501100011033 and WeLee project (1380939, FEDER Andalucía 2014-2020) funded by the Andalusian Regional Government. Salud María Jiménez-Zafra has been partially supported by a grant from Fondo Social Europeo and the Administration of the Junta de Andalucía (DOC_01073), and the article 83 contract with the company OpenSpring IT IBERIA S.L. (EXP. 2022_161). Thanks to SoGoodData for their support and for making the creation of this team a reality.

## References

[1] Institute of Health Metrics and Evaluation. Global Health Data Exchange (GHDx), https://vizhub.healthdata.org/gbd-results, 2023. Accessed: 2023-03-04.

[2] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of eRisk at CLEF 2021: Early Risk Prediction on the Internet (Extended Overview), CLEF (Working Notes) (2021) 864–887.

[3] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of eRisk 2022: Early risk prediction on the Internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, Proceedings, Springer, 2022, pp. 233–256.

[4] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, eRisk 2023: Depression, Pathological Gambling, and Eating Disorder Challenges, in: Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III, Springer, 2023, pp. 585–592.

[5] S. Kayalvizhi, T. Durairaj, B. R. Chakravarthi, et al., Findings of the Shared Task on Detecting Signs of Depression from Social Media, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, 2022, pp. 331–338.

[6] A. M. Mármol-Romero, A. Moreno-Muñoz, F. M. Plaza-del-Arco, M. D. Molina-González, M. T. Martín-Valdivia, L. A. Ureña-López, A. Montejo-Ráez, Overview of MentalriskES at IberLEF 2023: Early Detection of Mental Disorders Risk in Spanish, Procesamiento del Lenguaje Natural 71 (2023).

[7] S. M. Jiménez-Zafra, F. Rangel, M. Montes-y Gómez, Overview of IberLEF 2023: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), co-located with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), CEUR-WS.org, 2023.

[8] A.-M. Bucur, I. R. Podină, L. P. Dinu, A psychologically informed part-of-speech analysis of depression in social media, arXiv preprint arXiv:2108.00279 (2021).

[9] S. Rude, E.-M. Gortner, J. Pennebaker, Language use of depressed and depression-vulnerable college students, Cognition & Emotion 18 (2004) 1121–1133.
[10] N. Ramirez-Esparza, C. Chung, E. Kacewic, J. Pennebaker, The psychology of word use in depression forums in English and in Spanish: Testing two text analytic approaches, in: Proceedings of the international AAAI conference on web and social media, volume 2, 2008, pp. 102–108.
[11] T. Nguyen, A. Yates, A. Zirikly, B. Desmet, A. Cohan, Improving the generalizability of depression detection by leveraging clinical questionnaires, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 8446–8459. URL: https://aclanthology.org/2022.acl-long.578. doi:10.18653/v1/2022.acl-long.578.
[12] M. De Choudhury, M. Gamon, S. Counts, E. Horvitz, Predicting depression via social media, in: Proceedings of the international AAAI conference on web and social media, volume 7, 2013, pp. 128–137.
[13] D. M. Low, L. Rumker, T. Talkar, J. Torous, G. Cecchi, S. S. Ghosh, Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study, Journal of medical Internet research 22 (2020) e22635.
[14] R. Salas-Zárate, G. Alor-Hernández, M. d. P. Salas-Zárate, M. A. Paredes-Valverde, M. Bustos-López, J. L. Sánchez-Cervantes, Detecting depression signs on social media: a systematic literature review, in: Healthcare, volume 10, MDPI, 2022, p. 291.
[15] T. Zhang, A. M. Schoene, S. Ji, S. Ananiadou, Natural language processing applied to mental illness detection: a narrative review, NPJ digital medicine 5 (2022) 46.
[16] S. R. Kroenke K, W. JB, The PHQ-9: validity of a brief depression severity measure, J Gen Intern Med 16 (2001) 606–613.
[17] A. P. Association, The Diagnostic and Statistical Manual of Mental Disorders (5th ed.), 2013.
[18] S. Loria, Textblob Documentation, Release 0.16 (2020).
[19] C. Hutto, E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, in: Proceedings of the international AAAI conference on web and social media, volume 8, 2014, pp. 216–225.
[20] L. Zhuang, L. Wayne, S. Ya, Z. Jun, A robustly optimized BERT pre-training approach with post-training, in: Proceedings of the 20th Chinese National Conference on Computational Linguistics, Chinese Information Processing Society of China, Huhhot, China, 2021, pp. 1218–1227. URL: https://aclanthology.org/2021.ccl-1.108.
[21] P. Kralj Novak, J. Smailović, B. Sluban, I. Mozetič, Emoji sentiment ranking 1.0 (2015).
[22] L. Hanu, Unitary team, Detoxify, Github. https://github.com/unitaryai/detoxify, 2020.
[23] S. M. Mohammad, P. D. Turney, Crowdsourcing a word–emotion association lexicon, Computational intelligence 29 (2013) 436–465.
[24] S. Bird, NLTK: the natural language toolkit, in: Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, 2006, pp. 69–72.
[25] L. Feng, M. Jansche, M. Huenerfauth, N. Elhadad, A comparison of features for automatic readability assessment (2010).
[26] J. P. Kincaid, L. J. Delionbach, Validation of the automated readability index: A follow-up,

Human Factors 15 (1973) 17–20.
[27] G. Owen, Characterization of the Banzhaf–Coleman index, SIAM Journal on Applied Mathematics 35 (1978) 315–327.
[28] R. Gunning, The fog index after twenty years, Journal of Business Communication 6 (1969) 3–13.
[29] J. Anderson, Lix and rix: Variations on a little-known readability index, Journal of Reading 26 (1983) 490–496.
[30] K. L. Grabeel, J. Russomanno, S. Oelschlegel, E. Tester, R. E. Heidel, Computerized versus hand-scored health literacy tools: a comparison of Simple Measure of Gobbledygook (SMOG) and Flesch-Kincaid in printed patient education materials, Journal of the Medical Library Association: JMLA 106 (2018) 38.
[31] G. R. Klare, A table for rapid determination of Dale-Chall readability scores, Educational Research Bulletin (1952) 43–47.
[32] M. Ali, PyCaret: An open source, low-code machine learning library in Python, 2020. URL: https://www.pycaret.org, pyCaret version 1.0.0.
[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.