UC3Mental at MentalRiskES 2025: RF-SVM Ensemble Approach for Early Detection of Mental Health Risks Using NLP

Maximo Rodriguez[1,†], Pablo Zubasti[1,†] and Mario Saiz[1,†]

[1]*Universidad Carlos III de Madrid, 28911, Leganés, Madrid, Spain*

Abstract

This paper presents the methodologies employed by the UC3Mental team in its participation in the MentalRiskES tasks at IberLEF 2025. The competition focused on two key challenges related to the detection of mental illness in Spanish-language social media: addiction detection and addiction type classification. Our approach involved three different strategies: (1) a baseline model using a Support Vector Machine (SVM); (2) a two-stage pipeline that first applied a Random Forest classifier to classify addiction type, followed by a specialized SVM trained to detect addiction cases of that specific type; and (3) a model based on a BERT Transformer architecture.

Keywords

MentalRiskES, SVM, Random Forest, BERT, Natural Language Processing

1. Introduction

Mental disorders are commonly associated with significant disturbances in an individual's thoughts, emotions, or behavior. Although awareness of mental health issues has grown, the number of people affected continues to rise, and many still face stigma and insufficient access to treatment. One of the main challenges lies in the early detection of such conditions, which is often hindered by a lack of resources and timely intervention.

In recent years, the analysis of social media content has gained prominence as a complementary method for identifying mental health risks [1, 2, 3, 4, 5]. Social platforms offer a vast and accessible source of user-generated data that can be leveraged to detect potential signs of mental disorders [6, 7, 8, 9]. Nevertheless, this task presents several challenges, including the limited availability of annotated datasets, variability in language use, and the need for models capable of handling informal and diverse text.

MentalRiskES competition was designed to support research in this area, putting extra value on the early detection and classification of addiction-related not only on accurate results. The 2025 edition of the competition comprised two tasks: addiction detection and addiction type classification [10, 11]. The UC3Mental team participated in both tasks and explored three different approaches to address them:

- **Support Vector Machine (SVM):** A traditional machine learning method used as a baseline. This model was trained on text features to perform binary classification for addiction detection.
- **Random Forest + SVM Pipeline:** A two-step approach where a Random Forest model was first used to detect the type of addiction (topic classification). Based on the output, a specialized SVM was then selected and applied to determine the presence of addiction.
- **BERT-based Transformer:** A modern deep learning approach using a pre-trained BERT model fine-tuned on the competition dataset.

IberLEF 2025, September 2025, Zaragoza, Spain

[†]These authors contributed equally.

maxrodri@inf.uc3m.es (M. Rodriguez); pzubasti@pa.uc3m.es (P. Zubasti); masaizf@pa.uc3m.es (M. Saiz)

https://github.com/MaximoRdz/NLP-MENTALRISK-IBERLEF-2025 (M. Rodriguez)

0009-0005-2346-1429 (M. Rodriguez); 0009-0006-1906-118X (P. Zubasti)

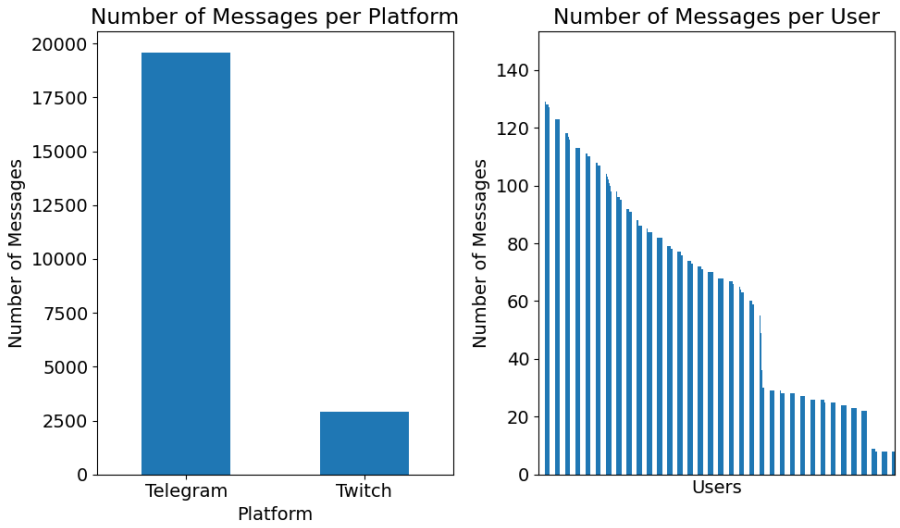


Figure 1: Distribution of messages across different platforms and users.

The objective was to assess the effectiveness of combining classical machine learning techniques with more recent deep learning architectures for the task of mental health risk detection on social media platforms.

Task Description

Task 1: Addiction Risk Detection

This is a binary classification task aimed at determining whether a user is at high risk (label = 1) or low risk (label = 0) of developing a gambling-related disorder based on their messages. The objective is to enable early detection and facilitate timely interventions.

Task 2: Type of Addiction Detection

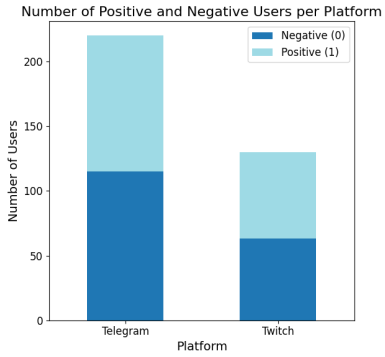
In this task, all users are considered to be at some level of risk (either low or high). Regardless of the risk level, the model must identify the specific type of addiction associated with the user's messages. The available labels for classification are: *Betting*, *Online Gaming*, *Trading*, and *Lootboxes*.

2. Exploratory Data Analysis

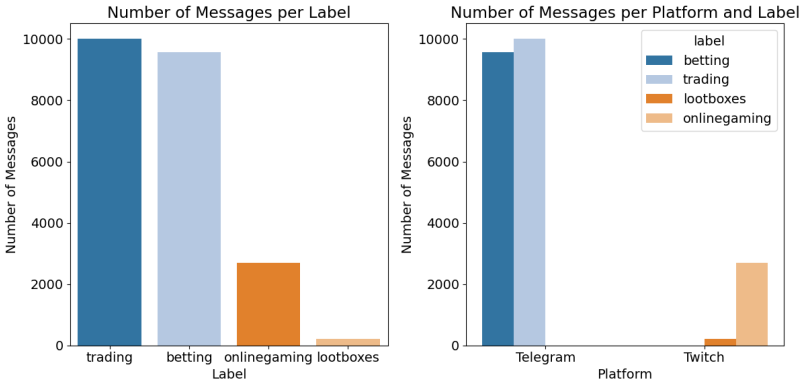
The dataset provided for the shared tasks consists of 358 annotated user samples for the addiction risk detection task (Task 1). The same set of users is used for the addiction type classification task (Task 2), but annotated with one of four specific addiction categories [12].

The dataset contains user-generated messages sourced from two platforms: Telegram and Twitch. Each user has a varying number of messages, and these messages serve as input features for both tasks. The dataset offers a realistic setting but presents some challenges related to class and platform imbalance, which are critical to consider during preprocessing and model development.

The initial analysis of the dataset reveals that the binary classification task (Task 1) is well-balanced, with an approximately equal distribution of high-risk and low-risk users. However, a significant platform imbalance is present: Telegram contributes a much larger volume of messages than Twitch. Despite this imbalance, both platforms include users from both risk categories. Additionally, users vary considerably in their message activity, which can be grouped into three general tiers: those with fewer than 20 messages, between 20 and 60 messages, and more than 60 messages, Figure 1. This variability should be considered when designing input representations or aggregating features across user histories.

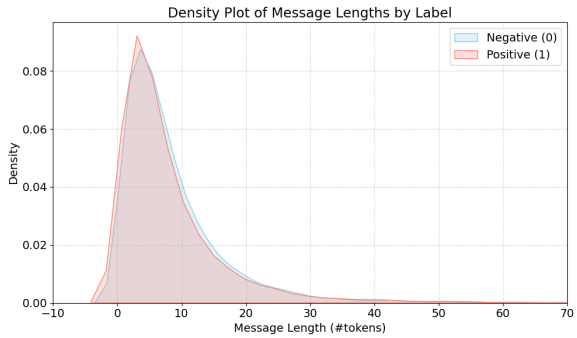


(a) Distribution of high-risk and low-risk users across platforms.



(b) Message length distribution by risk label.

Figure 2: Exploration of Task 1 data: risk label and text length characteristics.



(a) Number of messages per user by addiction type label.

(b) Message length distribution by addiction class.

Figure 3: Exploration of Task 2 data: addiction type distribution and content length.

In the multi-class classification task (Task 2), we observe a pronounced class imbalance, Figure 2b. The most frequent category is *Trading*, followed by *Online Gaming* and *Betting*, with *Lootboxes* being the least represented. Furthermore, there is a strong association between platform and addiction type: Telegram is the predominant source for messages related to *Betting* and *Trading*, while *Online Gaming* and *Lootboxes* are more commonly found on Twitch. Although these patterns are informative, caution must be exercised to avoid developing models that rely on superficial cues such as the platform itself. Overfitting to platform-specific distributions can hinder generalization and reduce the model's ability to focus on meaningful linguistic and behavioral features indicative of each addiction type. Finally, the hypothesis regarding a possible correlation between message length and addiction classification was addressed by examining the message length distribution across the full dataset, both for low and high risk labels and for addiction classification, see Figure 3a and 3b, respectively. Any exploitable correlation was discarded as no statistically meaningful difference could be appreciated on the distributions.

In addition to the descriptive analysis of the data, and as a consequence of the poor results obtained in Task 1 during the training and testing of the models (to be discussed later), a further investigation was conducted to assess the hypothesis regarding the absence of an underlying pattern governing the message labeling process. Specifically, the study aimed to validate the notion that the target variable for Task 1 is distributed in an almost random manner, which would mathematically preclude the development of models capable of achieving significantly higher accuracy than 50%. The following set of experiments was conducted to attempt to validate the previously stated hypothesis:

1. Train three basic machine learning models and evaluate their performance on the test set. If different techniques yield results close to 50% accuracy, this could suggest a random distribution of

the labels. However, such results alone are not sufficient to conclusively support this hypothesis.

2. Apply a k-means clustering model to the vectorized data (using TF-IDF, as detailed later in the document), where each vector represents a user. The goal is to group similar vectors—hence, similar users—and subsequently analyze the statistical distribution of Task 1 labels within each cluster. The underlying assumption is that similar users should exhibit similar labels. If the label frequencies are comparable across clusters, this could indicate the absence of a clear labeling pattern.
3. Perform a Kullback-Leibler (KL) divergence test to compare the "distance" between the empirical probability distribution of the target variable in Task 1 and a uniform binary probability distribution (i.e., a Bernoulli distribution with 50% probability for each label).

Training of three basic ML models

The three models employed were: k-Nearest Neighbors (k-NN), logistic regression, and a decision tree (ID3). The results in terms of classification accuracy and Area Under the ROC Curve (AUC) are presented in Table 1.

Machine Learning model	Accuracy	f1-score (macro-avg)	f1-score (weighted-avg)	AUC
k-NN	51.38%	0.34	0.35	0.58
Logistic Regression	70.83%	0.71	0.71	0.77
Decision Tree	58.33%	0.58	0.58	0.59

Table 1
Classification results of the three basic ML models employed for the randomness test of the objective variable in task 1.

Based on the results, two of the three models exhibit values close to 50%, while logistic regression approaches 70%. These figures should be interpreted with caution, as although most models appear to be operating on a randomly labeled target variable, the notably higher performance of logistic regression may be a consequence of the specific random train-test split used. This could have fortuitously resulted in overly optimistic test performance that is not generalizable.

Clustering of similar users

In this section, as previously mentioned, the aim is to group similar users based on the TF-IDF vectors generated during the text processing phase (further details are provided in the following sections of this article). The k-means algorithm was employed for clustering the users, and the optimal number of clusters k was determined using the Silhouette Score (Equation 1). This metric, applied by performing multiple clustering runs with varying k values, enables the identification of the most appropriate number of clusters. By applying the algorithm and the silhouette criterion, the optimal number of clusters was found to be 2 (see Figure 4).

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{1}$$

Where $a(i)$ is the average distance between point i and all other points in the same cluster and $b(i)$ is the average distance between point i and all points in the nearest cluster to which i does not belong. The average silhouette value over all points provides an overall measure of clustering quality.

By characterizing each of the resulting groups, we obtain the results shown in Figure 5.

From this, it can be concluded that no clear pattern exists namely, that one type of user (i.e., one of the two clusters) has significantly more labels of one class than the other suggesting that identical or highly similar inputs yield completely different outputs.

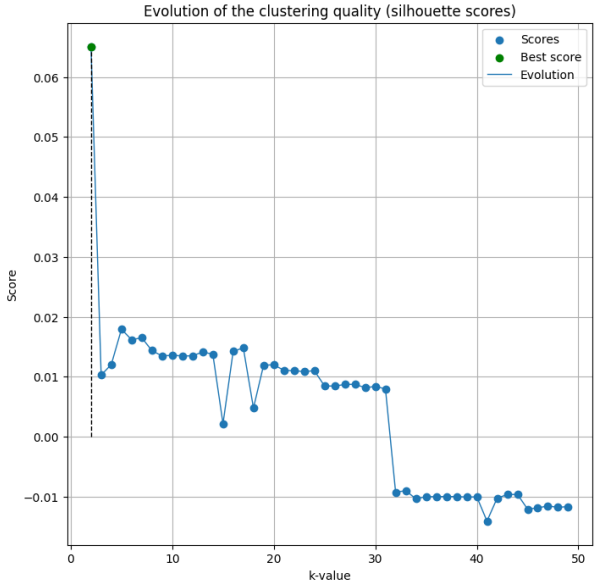


Figure 4: Evolution of the Silhouette Score across different values of k. Higher scores indicate better clustering quality.

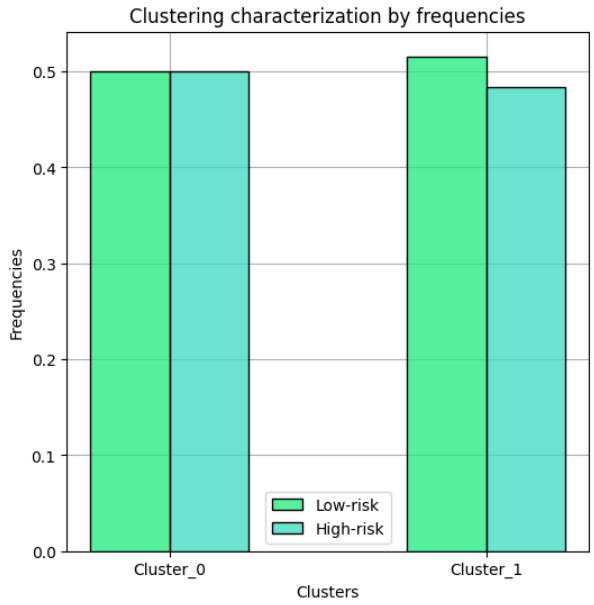


Figure 5: Characterization performed on each of the two identified clusters. As can be observed, similar vectors (i.e., similar users) within the same cluster are being labeled as 0 or 1 in an almost perfectly equiprobable manner.

KL-divergence test

The Kullback-Leibler (KL) divergence is a measure of how one probability distribution differs from a second, reference probability distribution. It is important to notice that KL divergence is not symmetric, meaning that $D_{\mathrm{KL}}(P \parallel Q) \neq D_{\mathrm{KL}}(Q \parallel P)$. So it is very important to note that KL divergence is not a distance, but rather a similarity measure.

For continuous probability density functions $p(x)$ and $q(x)$, the KL divergence is defined as:

$$D_{\mathrm{KL}}(p \parallel q) = \int_{-\infty}^{\infty} p(x) \log \left(\frac{p(x)}{q(x)} \right) dx \tag{2}$$

In both cases, the KL divergence is always non-negative and equals zero if and only if $P = Q$ (or

$p(x) = q(x)$ almost everywhere in the continuous case).

For two discrete probability distributions P and Q defined over the same set $\mathcal{X}$, the KL divergence from Q to P is given by:

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right) \tag{3}$$

To obtain a unified KL divergence result, the arithmetic mean of the KL divergence calculated in both directions is applied, as shown in Equation 4.

$$D^*_{\mathrm{KL}}(P \parallel Q) = \frac{1}{2} \left[D_{\mathrm{KL}}(P \parallel Q) + D_{\mathrm{KL}}(Q \parallel P) \right] \tag{4}$$

As a result of applying Equation 4, the bidirectional KL divergence is found to be 0.0001922, a value sufficiently small and close to zero to support the assumption that the probability distribution of the target variable in Task 1 closely resembles a uniform distribution (i.e., a random distribution). Therefore, based on the results obtained from the preceding experiments, we can reasonably conclude that the target variable in Task 1 has been labeled in a random manner, or at the very least, lacks an underlying pattern that can be effectively learned by machine learning models.

3. Approaches and contribution

The following section presents a detailed account of the proposed systems and models employed for the text classification task, ranging from traditional baseline approaches, through their combinations, and culminating in transformer-based methods and large language models (LLMs).

3.1. SVM baseline

The initial approach employed to address the task of user classification based on their messages was a traditional method relying on Support Vector Machines (SVMs), adopted as a baseline model. This model takes as input the users' pre-processed texts, as illustrated in Figure 6. The rationale behind this approach is to assess the effectiveness of a classical method before resorting to large language models and transformer-based architectures, given that such traditional algorithms have demonstrated strong performance in past competitions.

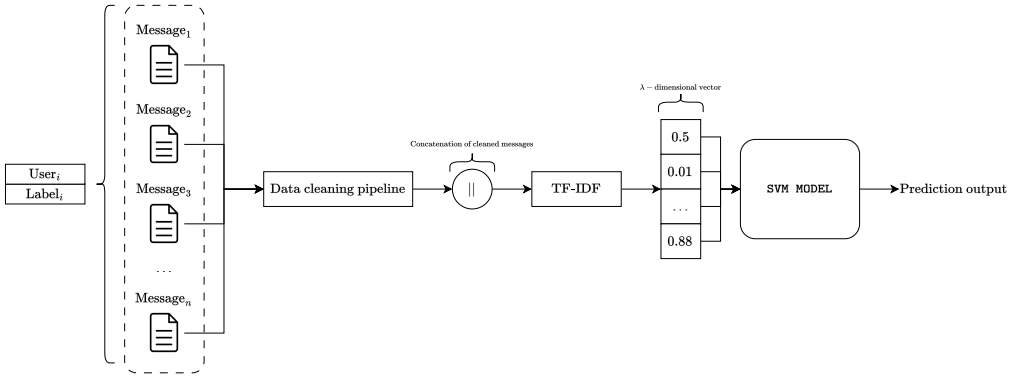


Figure 6: Overview of the user text reprocessing pipeline. The procedure includes a message cleaning phase, which involves the removal of emojis (optionally, at the user's discretion), the elimination of punctuation marks, and the use of *TweetTokenizer* for tokenization.

The main aspect of the classical SVM-based approach lies in the preprocessing strategy chosen to clean and vectorize the texts in a manner suitable for the SVM model to process the information. As illustrated in Figure 6, the core idea is to tokenize the texts using *TweetTokenizer*, optionally remove emojis, and eliminate punctuation marks. It is important to note that, despite efforts to thoroughly

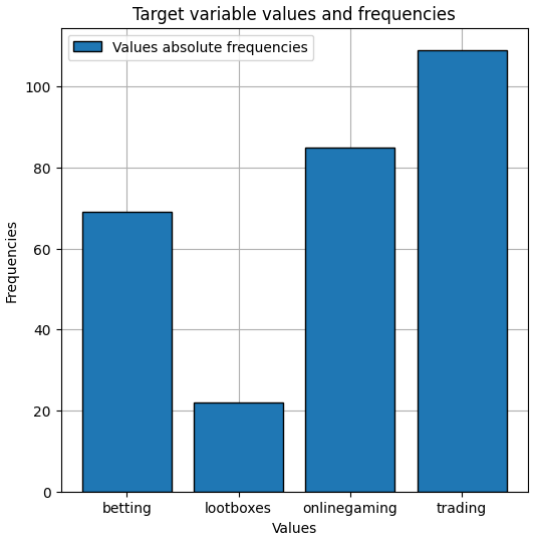


Figure 7: Distribution of the imbalanced classes for Task 2.

clean the texts during preprocessing, the data originates from social media conversations. As such, spelling errors and informal or inconsistent semantic and syntactic structures present in the messages significantly hinder the ability to obtain clean, easily processable text—this being one of the major challenges of the task.

Since both tasks involve classifying users based on their texts, all available messages per user were tokenized, cleaned, and concatenated, such that the input for each user corresponds to the concatenation (with whitespace inserted between them) of all their texts (see Equation 5). This input is then processed by the TF-IDF algorithm, which is used to vectorize the information that will be provided to the SVM.

$$User_i \leftarrow \Big\|_{msg \in \mathcal{M}_i} \Phi(msg) \;||\; \lambda;\; \forall i \in |Users| \tag{5}$$

Where $\mathcal{M}$ denotes the set of messages associated with the i-th user, Φ represents the tokenization and data cleaning function, and λ denotes the whitespace character.

For the training and hyperparameter tuning of the SVM model, an 80–20 split was performed for the training and test sets, respectively, selected at random while ensuring a stratified distribution across all classes. In the case of Task 1, the label distribution between high_risk (1) and low_risk (0) was nearly balanced at 50%, thus eliminating the need for additional class balancing techniques. However, for Task 2, the multiclass classification problem exhibited the class imbalance shown in Figure 7, prompting the use of the SMOTE technique to perform oversampling and achieve a more balanced class distribution. SMOTE was applied exclusively to the training set, as applying it to the entire dataset (including the test set) would result in what is known as data leakage—an issue that introduces unintended knowledge into the model and leads to overly optimistic performance results.

Hyperparameter	Values to test
C	0.1, 1, 2, 3, 7
γ	`scale`, 0.01, 0.001, 0.0001
Kernel function	`linear`, `rbf`

Table 2
Hyperparameter search space for the SVM baseline model.

Finally, for hyperparameter tuning (conducted via grid search), various values of C, γ, and the SVM kernel were tested (see Table 2), using 5-fold cross-validation. The optimal configuration was found to be: C = 3, γ = auto-scaled, and kernel = radial basis function (RBF).

3.2. RF-SVM ensemble

The RF-SVM ensemble constitutes the main contribution of the present work. The rationale behind this ad hoc model, specifically designed to address the early detection task, lies in the observation that both Task 1 and Task 2 are interrelated, as they rely on the same underlying data (i.e., the same users and messages), differing only in the type of labels provided. While the SVM model yielded the best performance for Task 1, for Task 2, the Random Forest model slightly outperformed the SVM, albeit by a narrow margin. Accordingly, the RF-SVM ensemble operates as follows (see Figure 8): a Random Forest model, trained for Task 2, is first employed to classify the message, achieving excellent precision, as will be discussed later. Based on the classification output for Task 2, a second classification is then performed using an SVM model that has been trained exclusively on messages originally labeled with the corresponding addiction type. In this manner, a general Random Forest classifier determines the type of addiction (among the four available classes) to which the current message belongs, allowing a specialized SVM—trained specifically on messages related to that addiction type—to carry out Task 1, which is inherently more challenging.

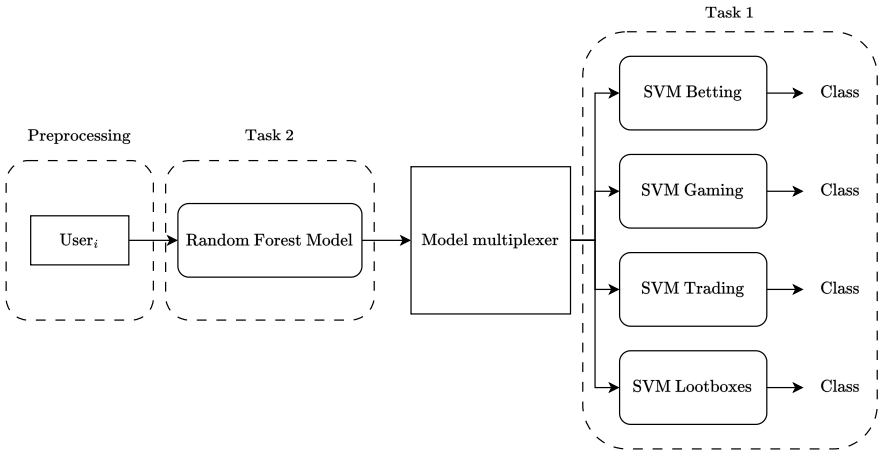


Figure 8: Architecture of the RF-SVM ensemble system. A Random Forest model solves Task 2 in order to partition the problem into a smaller classification subtask (Task 1), which is addressed by a dedicated SVM model.

This strategy aims to enable model training at a finer level of granularity, where, instead of learning all distinctions and characteristics across the entire dataset, the model captures local patterns. For example, it may be significantly more common to observe high-risk mental health indicators in messages associated with a specific type of addiction than in others.

3.3. BERT

BERT (Bidirectional Encoder Representations from Transformers) is a language model developed by Google in 2018. Its primary innovation lies in the use of a bidirectional Transformer architecture, which enables the model to understand the context of a word by considering both the preceding and succeeding words. This capability makes it particularly effective for tasks such as sentiment analysis, question answering, and text classification. Within the scope of the problem to be addressed (Task 1 and Task 2), the Transformer-based approach aimed to strike a balance between result quality and the use of models that do not impose excessive computational or memory costs. To meet both requirements, BERT-small [13] was employed. BERT-small is a reduced version of BERT designed to be more resource-efficient. While the base BERT model consists of 12 Transformer layers and approximately 110 million parameters, BERT-small typically includes only 4 layers and around 29 million parameters, resulting in a faster and more lightweight model. Although it trades off some accuracy compared to its larger counterparts, BERT-small is well-suited for deployment on resource-constrained environments such as mobile devices or embedded systems.

The data preprocessing followed the previously described pipeline, with the exception of the steps beginning from the application of the TF-IDF algorithm (inclusive), as the Transformer architecture inherently computes its own embeddings. In other words, external vectorization of the texts is not required. The procedure essentially consists of performing fine-tuning on the pre-trained BERT-tiny base model in order to adapt it to the desired text classification task. Within the fine-tuning process, the following selected parameters are of particular note:

- **Evaluation strategy**: per epoch
- **Train batch size**: 4
- **Evaluation batch size**: 4
- **Training epochs**: 25
- **Learning rate**: $7 \cdot 10^{-7}$
- **Weight decay**: 10^{-3}
- **Model storage policy**: best model

3.4. LLM

Additionally, although it was not part of the models used in the competition, an approach based on Large Language Models (LLMs) was tested to address the classification task. The large language model (LLaMA 3, 70B) [14] developed by Meta AI was used to perform two classification tasks on user-generated text. The LLaMA 3 model is a transformer-based architecture trained on a diverse and extensive corpus of publicly available web data, books, academic texts, and code. It is optimized for instruction following, question answering, and general-purpose language understanding tasks. The 70-billion parameter version of the model was used, which represents the most capable publicly released version of the LLaMA 3 family at the time of this study. All classification tasks were carried out using a zero-shot prompting setup — no additional fine-tuning or task-specific training data was provided to the model. Instead, the model was prompted with natural language instructions tailored to each task and expected to infer the appropriate label based on its pre-trained knowledge and language understanding capabilities. The results of the LLM-based approach were not successful in either of the two tasks, as will be shown in the following section on experimental results. Therefore, as previously mentioned, only the first three approaches described in this section were employed for the competition. To conclude this section, the prompts used to address Task 1 and Task 2 are provided below:

Prompt for task 1

```
You will now be given a series of messages written by a person. Your task
is to analyze the content and determine if there are clear and repeated
indications of behaviors related to pathological gambling (gambling addiction).
Frequent references to gambling or compulsive trading. Expressions of loss of
control, anxiety about gambling, or urgent need to win back money. Mentions
of significant financial losses due to these games or similar behaviors.
Difficulty to stop gambling or thinking about gambling constantly. You should
not interpret all comments about gambling or trading as a sign of addiction
since the objective is only to detect cases of pathological gambling, addiction
and loss of control. Your answer must be exclusively a number: 1 if there are
several clear and direct indications of pathological gambling. 0 if there is
not enough clear evidence (even if there are vague or isolated mentions). Now
evaluate the following messages:
```

Prompt for task 2

```
Below you will be given a series of messages written by a person suffering
from some form of pathological gambling (gambling addiction). The goal is to
```

```
classify the condition according to the specific type of addictive behavior.
The following categories exist: Betting: refers to addiction to traditional
betting, such as sports betting, horse racing or gambling in physical or
online casinos (roulette, poker, etc.). Onlinegaming: Refers to addiction to
multiplayer online video games, where the person plays compulsively without
there necessarily being a direct monetary transaction. Lootboxes: Addiction
to random reward systems within video games, where a "box" is bought with
real money to obtain virtual items, generating a behavior similar to that of
slot machines. Trading: encompasses the addiction to trading financial assets
(such as cryptocurrencies, stocks or forex), in which the individual performs
compulsive operations seeking immediate gratification, assuming high risks.
Your answer should be exclusively one of these words, with no additional
phrases, no quotation marks and no explanations: betting, onlinegaming,
lootboxes, trading. Now evaluate according to the following messages:
```

4. Experimental results

4.1. Task 1: Gambling Disorder Detection

As discussed before, the goal of the first task was to identify whether a user's messages indicated signs of a gambling disorder. Each user was labeled using a binary classification:

- **0** — No indication of a gambling disorder.
- **1** — Language indicating the presence of a gambling disorder.

The results obtained on the test set from the trained models, following hyperparameter tuning, are presented in Table 3.

Approach	Accuracy	f1-score (macro-avg)	f1-score (weighted-avg)	Precision	Recall
SVM baseline	62%	0.62	0.62	0.63	0.63
RF-SVM ensemble	69.50%	0.685	0.69	0.7075	0.69
BERT-small	59.72%	-	-	-	-
LLaMA 3	54.9%	0.51	0.52	0.58	0.28

Table 3
Results obtained during Task 1.

As shown in Table 3, the best-performing approach for Task 1 was the proposed RF-SVM ensemble, achieving approximately 70% accuracy on the test set. However, these results proved to be optimistic when compared to those obtained during the competition, where the accuracy was around 52%.

4.2. Task 2: Gambling Behavior Typology

The second task aimed to further categorize the type of gambling-related behavior expressed by the user. The model was prompted to classify the content of each user's messages into one of four predefined categories, reflecting distinct behavioral patterns:

- **betting** — Traditional gambling (e.g., sports betting, horse racing, casinos).
- **lootboxes** — Randomized virtual item purchases in video games.
- **onlinegaming** — Compulsive play of online games without monetary stakes.
- **trading** — High-risk trading of financial assets such as cryptocurrency or stocks.

The results obtained on the test set from the trained models, following hyperparameter tuning, are presented in Table 4.

Approach	Accuracy	f1-score (macro-avg)	f1-score (weighted-avg)	Precision	Recall
SVM baseline	94%	0.81	0.93	0.96	0.80
RF-SVM ensemble	94%	0.81	0.93	0.96	0.80
BERT-small	97.22%	-	-	-	-
LLaMA 3	68.60%	0.58	0.63	0.73	0.31

Table 4
Results obtained during Task 2.

As shown in Table 4, the first notable observation is that the results for Task 2 from the SVM baseline model and the RF-SVM ensemble are identical. This has a straightforward explanation: recalling the structure of the RF-SVM approach, a Random Forest model is first used to solve Task 2, and its output is then passed to a dedicated SVM to resolve Task 1. In other words, the ensemble's contribution is specific to Task 1. Since Task 2 is resolved by a Random Forest model (whose results on the test set happen to match those of the SVM baseline), the RF-SVM ensemble behaves identically to the SVM baseline for Task 2.

In terms of accuracy, the BERT-small-based approach stands out by achieving 97.22%, an excellent result—though not dramatically higher than those achieved by the SVM and RF-SVM models. It is important to note that comparable results to those obtained with BERT-small were approximately replicated using machine learning models that are significantly less costly in terms of computational time and memory, reinforcing the notion that larger and more modern models are not always the best solution.

Finally, the LLM-based approach performed noticeably worse. The results obtained with LLaMA 3 were significantly inferior to those of the other models. This is particularly striking given that, unlike Task 1, Task 2 appears to be a well-defined task that can be effectively solved, with most models achieving classification metrics well above 90%.

5. Conclusions

This article has presented an innovative ensemble architecture that combines advanced machine learning models to address both binary and multiclass text classification tasks. The results obtained on the test sets demonstrate the improvements brought by this new approach, enabling the achievement of eighth place in the MentalRiskES competition, held as part of IberLEF 2025. Performance on Task 1 was significantly lower than on Task 2, as discussed in the article. The label distribution for Task 1 appears to follow a random pattern, lacking identifiable structures or features that machine learning models can effectively learn. In comparison with modern approaches based on Transformers and Large Language Models (LLMs), our ensemble method performs equally well or better, while offering the added benefit of reduced computational cost in terms of both time and memory. This efficiency was further reflected in the system's ranking as the second fastest among all submitted solutions for the proposed tasks.

6. Future work

As future work, we propose a more extensive study and experimentation using a larger dataset (with more instances), in which the labeling of messages and users has been thoroughly reviewed and there is evidence supporting the existence of a modelable phenomenon that can be learned through machine learning techniques. The RF-SVM architecture is not limited exclusively to the tasks proposed in the MentalRiskES competition; rather, it can be extended through basic modifications to multiclass classification problems where the data is doubly labeled (in binary or multiclass form), enabling predictions on data subsets using models with lower granularity.

7. Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT to assist with English translation, grammar correction, and spelling checks. All content generated with the help of this tool was reviewed and edited by the author(s), who take full responsibility for the final version of the publication.

References

[1] S. Chancellor, M. D. Choudhury, Methods in predictive techniques for mental health status on social media: a critical review, NPJ Digital Medicine 3 (2020) 43. URL: https://www.nature.com/articles/s41746-020-0233-7. doi:10.1038/s41746-020-0233-7, © The Author(s) 2020. eCollection 2020.

[2] R. A. Calvo, D. N. Milne, M. S. Hussain, H. Christensen, Natural language processing in mental health applications using non-clinical texts, Natural Language Engineering 23 (2017) 649–685. doi:10.1017/S1351324916000383.

[3] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, J. C. Eichstaedt, Detecting depression and mental illness on social media: an integrative review, Current Opinion in Behavioral Sciences 18 (2017) 43–49. URL: https://www.sciencedirect.com/science/article/pii/S2352154617300384. doi:https://doi.org/10.1016/j.cobeha.2017.07.005, big data in the behavioural sciences.

[4] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses, in: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Association for Computational Linguistics, Denver, Colorado, 2015, pp. 1–10. URL: https://aclanthology.org/W15-1201/. doi:10.3115/v1/W15-1201.

[5] A. Benton, M. Mitchell, D. Hovy, Multi-task learning for mental health using social media text, CoRR abs/1712.03538 (2017). URL: http://arxiv.org/abs/1712.03538. arXiv:1712.03538.

[6] D. E. Losada, F. Crestani, A test collection for research on depression and language use, in: N. Fuhr, P. Quaresma, T. Gonçalves, B. Larsen, K. Balog, C. Macdonald, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2016, pp. 28–39.

[7] A. Yates, A. Cohan, N. Goharian, Depression and self-harm risk assessment in online forums, in: M. Palmer, R. Hwa, S. Riedel (Eds.), Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2968–2978. URL: https://aclanthology.org/D17-1322/. doi:10.18653/v1/D17-1322.

[8] P. Resnik, W. Armstrong, L. Claudino, T. Nguyen, V.-A. Nguyen, J. Boyd-Graber, Beyond LDA: Exploring supervised topic modeling for depression-related language in Twitter, in: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Association for Computational Linguistics, Denver, Colorado, 2015, pp. 99–107. URL: https://aclanthology.org/W15-1212/. doi:10.3115/v1/W15-1212.

[9] K. Saha, J. Torous, S. K. Ernala, C. Rizuto, A. Stafford, M. Choudhury, A computational study of mental health awareness campaigns on social media, Translational Behavioral Medicine 9 (2019). doi:10.1093/tbm/ibz028.

[10] J. A. González-Barba, L. Chiruzzo, S. M. Jiménez-Zafra, Overview of iberlef 2025: Natural language processing challenges for spanish and other iberian languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS.org, 2025.

[11] A. M. Mármol-Romero, P. Álvarez Ojeda, A. Moreno-Muñoz, F. M. P. del Arco, M. D. Molina-González, M.-T. Martín-Valdivia, L. A. Ureña-López, A. Montejo-Ráez, Overview of mentalriskes at iberlef 2025: Early detection of mental disorders risk in spanish, Procesamiento del Lenguaje Natural 75 (2025).

[12] P. Álvarez Ojeda, M. V. Cantero-Romero, A. Semikozova, A. Montejo-Ráez, The precom-sm

corpus: Gambling in spanish social media, in: Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 17–28.

[13] H. Tsai, J. Riesa, M. Johnson, N. Arivazhagan, X. Li, A. Archer, Small and practical BERT models for sequence labeling, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3632–3636. URL: https://aclanthology.org/D19-1374/. doi:10.18653/v1/D19-1374.

[14] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. URL: https://arxiv.org/abs/2302.13971. arXiv:2302.13971.