

Monitoring Therapeutic Plans and Risk Signals from Clinical Narratives in Mental Health using Natural Language Processing

Margarita Ruiz Olazar, Andrea Aguilera, Diego Ihara,
Fernando Sosa, Cecilia Scales, Benjamin Baran

Universidad Comunera, SmartDataLab
Asuncion, Paraguay
margarita.ruiz@ucom.edu.py

Abstract

Clinical narratives in electronic health records (EHRs) are a key source of information in mental health care, but their unstructured nature limits systematic analysis. We propose an NLP framework to extract therapeutic plan information and identify indirect clinical risk signals from psychiatric narratives in Spanish. Using a retrospective dataset of outpatient consultations, we develop an interpretable approach combining rule-based extraction with a risk model based on TF-IDF and logistic regression, incorporating probability calibration and asymmetric thresholds to prioritize high-risk cases. We compare this approach against a fine-tuned transformer model (RoBERTuito) using patient-level cross-validation. These findings demonstrate that combining interpretability, robust evaluation, and domain-specific considerations enables the development of practical NLP tools for mental health applications using real-world clinical data.

Keywords: Mental Health, Electronic Health Records, Natural Language Processing, Risk Identification.

Introduction

Clinical narratives recorded in electronic health records (EHRs) constitute a central source of information in mental health care, capturing symptoms, therapeutic decisions, and longitudinal patient evolution (World Health Organization 2025). However, their unstructured and highly contextual nature limits their systematic use for monitoring treatment trajectories and identifying patients at potential clinical risk. This challenge is particularly relevant in public mental health systems, where clinical documentation is predominantly narrative and the availability of structured, annotated datasets is limited. As a result, critical information remains embedded in free text, hindering large-scale analysis and timely identification of patients requiring closer clinical attention.

Recent advances in natural language processing (NLP) have enabled the extraction of clinically relevant information from unstructured text, including symptom detection, treatment analysis, and suicide risk assessment. However, many existing approaches rely on large-scale datasets,

opaque deep learning models, or direct prediction of clinical outcomes, raising concerns regarding interpretability, robustness, and safe deployment in real-world healthcare settings (Camacho-Collados et al. 2017)(Le Glaz et al. 2021).

In this work, we propose an interpretable NLP framework for (i) extracting structured therapeutic plan information and (ii) identifying indirect clinical risk signals from psychiatric narratives in Spanish. Importantly, we conceptualize risk as a latent construct inferred from linguistic patterns, rather than a direct diagnostic prediction. Specifically, we model clinical risk using sentiment-derived signals as a weak proxy for affective and cognitive states expressed in clinical narratives. Prior studies have shown that linguistic markers of negative affect, distress, and behavioral changes correlate with clinically relevant outcomes in mental health. In this context, sentiment is not treated as a diagnostic substitute, but as an auxiliary signal that contributes to identifying patients who may require further clinical attention.

Our approach combines rule-based extraction techniques, grounded in expert knowledge, with a calibrated machine learning model based on TF-IDF representations and logistic regression. To contextualize performance, we compare our approach against pre-trained and fine-tuned transformer-based models.

The main contributions of this work are:

- An interpretable pipeline for extracting structured therapeutic information from clinical narratives in Spanish.
- A clinically-aligned risk modeling approach that prioritizes sensitivity through probability calibration and asymmetric thresholds.
- An empirical evaluation demonstrating that simple and interpretable models can achieve competitive performance compared to transformer-based approaches in low-resource clinical settings.

This work contributes to the development of practical and responsible NLP tools for mental health, particularly in contexts where data scarcity and interpretability constraints are critical.

Material and Methods

This project was conducted at a specialized Mental Health Center that serves as the largest provider of mental health

services within the Paraguayan social security system. The institution’s catchment area spans the full socioeconomic spectrum of the country, ensuring broad population coverage. Its integrated electronic health record (EHR) system contains records for more than 1.4 million individuals, including detailed sociodemographic and clinical information, thereby constituting a large-scale and representative source of real-world mental health data.

This study employed an observational and retrospective design; no interventions or manipulations of variables were performed, and only previously recorded clinical data were used. The records analyzed correspond to outpatient psychiatric consultations conducted between 2021 and 2025. The cohort analyzed consists exclusively of patients diagnosed with depression, selected according to clinical criteria defined by their referring institution.

The primary unit of analysis is the clinical record from each consultation, although some analyses consider aggregating information at the patient level. The cohort size reflects a realistic clinical scenario in the public mental health system, where the availability of large volumes of structured, annotated data is often limited. Figure 1 summarizes the general processing flow of clinical text from its reception to the branching towards the specific analysis modules.

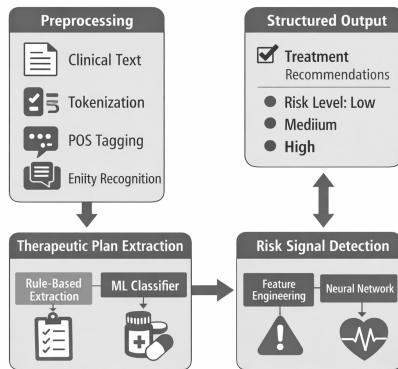


Figure 1: Interpretable NLP pipeline for therapeutic plan and risk signal extraction.

Corpus Understanding and Description

The corpus consists of 2,230 clinical reports from 59 anonymized patients diagnosed with depression and treated at a mental health institute in Asunción, enabling longitudinal analysis through multiple consultations per patient. The reports, written in Spanish by psychiatrists, are brief and reflect the concise style of outpatient clinical records.

The data are provided in CSV format, combining structured fields with narrative text, which facilitates computational processing. The corpus also includes a manual risk-level labeling system (Low, Medium, High), used as a reference to evaluate the developed model, without implying a clinical diagnosis.

Additionally, a dictionary of medications and clinical abbreviations validated by specialists was used to support therapeutic information extraction, reducing ambiguities and ensuring clinical coherence.

Anonymization and Data Protection Process Given the sensitive nature of the information analyzed, an irreversible anonymization process was implemented before the analysis began. This process included the removal of direct identifiers, such as first names, identification numbers, addresses, and any other data that could identify the patient or the healthcare professional. Once this process was completed, the data used did not allow for patient re-identification, thereby complying with the principles of confidentiality and personal data protection.

Data Preprocessing

A standardized cross-sectional preprocessing was applied to the clinical reports to ensure a homogeneous and reproducible foundation without compromising clinical integrity. Key steps included:

- **Data Selection:** Relevant features (*ID*, *Date*, *Report*, and *Risk*) were extracted from outpatient records of patients with depression.
- **Temporal Normalization:** The *consultation Date* was converted to a standard *datetime* format while maintaining traceability.
- **Text Optimization:** Clinical reports underwent normalization (encoding, case, and spacing corrections) and the removal of digitization artifacts.
- **Data Integrity:** Original records were preserved in parallel with the processed versions to allow for auditing and comparative analysis.

This standardized output serves as the common baseline for the specific transformations detailed in subsequent methodological chapters.

Corpus Limitations

While the corpus provides valuable real-world context, its findings are limited by a small sample size, variable text quality, and a lack of expert annotations. These factors restrict the study’s generalizability and highlight the need for further validation across different institutions and datasets.

However, the current corpus remains sufficient for testing the feasibility of the methodology and establishing a foundation for future research.

Preprocessing and normalization of clinical text

Each row of the dataset represents an individual consultation and includes basic structured information along with the free-text clinical account. Table 1 summarizes the columns selected for the analysis. The selection of these columns responds to the need to preserve both the temporal traceability of the queries and the textual information relevant to the extraction and modeling objectives.

Table 1: Columns of the Clinic Corpus used in the study

Column	Description
ID	Anonymized patient identifier
Date	Recorded date of the psychiatric consultation
Report	Narrative text of the consultation
Risk	Clinical risk label (Low, Medium, High)

A set of transformations applied to the field *Report*. The applied transformations are summarized in Table 2. Unlike aggressive text cleaning approaches, a conservative strategy was chosen, prioritizing the preservation of relevant semantic information. The resulting text was stored in a derived column, while the original report remained available for reference, comparison, and quality control. This decision allows for auditing the impact of preprocessing and prevents the irreversible loss of clinical information.

Table 2: Transformations Applied to Clinical Text

Stage	Description
Normalization	Unification of character encoding and removal of corrupt symbols
Correction of accents	Restoration of diacritics
Selective lowercase	To lowercase without altering clinical abbreviations
Cleaning of spaces	Removal of duplicate whitespace
Scanning errors	Removal of isolated characters sequences without semantic content

Therapeutic Plan Extraction

The extraction process focuses on identifying explicit information documented in the current consultation, particularly that associated with the section of the narrative corresponding to the therapeutic plan. The variables of interest are summarized below:

- **Medication:** Name of the drug prescribed in the consultation
- **Dose:** Amount of active ingredient prescribed
- **Posology:** Frequency and method of administration
- **Treatment schedule:** Timeline of treatment
- **Refill:** Indication for medication refill
- **Psychotherapy:** Referral to psychotherapy or psychoeducation
- **Next check-up:** Date or interval of the next follow-up

The approach adopted does not attempt to infer undocumented implicit information, but rather to faithfully structure the explicit data present in the clinical text. A rule-based approach using expert knowledge was chosen over deep learning models because the high variability of clinical text limits

the applicability of complex supervised approaches. Furthermore, the interpretability of the method is crucial in clinical contexts, where decision traceability is an ethical and methodological requirement. (Le Glaz et al. 2021)

Combining regular expressions with fuzzy matching techniques (Navarro 2001) allows for the capture of recurring linguistic patterns in clinical documentation, even in the presence of typographical errors or abbreviations, while maintaining explicit control over the extractor’s behavior.

Evaluation of the extractor The extractor was evaluated on a sample of manually reviewed clinical reports. The results show a high level of agreement in the identification of medications and associated attributes, confirming the viability of the proposed approach for structuring relevant therapeutic information in the mental health domain. This process was automated using Hugging Face spaces.

Risk Signal Detection

We model clinical risk as a three-class classification problem (Low, Medium, High), using manually annotated labels provided by a clinical expert. These labels are interpreted as a reference signal and not as diagnostic ground truth.

Sentiment as a Proxy Signal

In this work, sentiment is used as an indirect proxy for clinical risk. While sentiment does not directly represent clinical severity, it captures affective and cognitive signals embedded in clinical narratives, such as expressions of distress, hopelessness, or behavioral deterioration. Therefore, sentiment is treated as a weak signal that complements other sources of information in identifying patients who may require closer monitoring.

Feature Representation

Textual data were represented using TF-IDF features at both word and character levels. Specifically, we used unigrams and bigrams at the word level and character n-grams (n=3–5). The vocabulary was limited to the top 10,000 features based on term frequency, and terms with document frequency below 2 were excluded.

Model

We used a logistic regression classifier with L2 regularization (C=1.0) and class weighting to account for class imbalance. The probabilities generated by the best-performing pre-trained sentiment model were incorporated as additional features.

Data Splitting and Validation

To prevent information leakage, data were split at the patient level. We performed stratified 5-fold cross-validation, ensuring that all records from a given patient were contained within a single fold.

Experimental Setup and Baseline

To contextualize the performance of the proposed approach, we implemented a benchmarking framework:

1. **Transformer Baseline:** We implemented a transformer-based baseline using RoBERTuito (Pérez et al. 2022). The model was fine-tuned on the same dataset using a standard cross-entropy loss function. Fine-tuning was performed for 3 epochs with a learning rate of $2e-5$ and batch size of 16. Early stopping was applied based on validation loss.
2. **Optimized Model:** A composite classifier using TF-IDF (word/char n-grams) and Logistic Regression, augmented with the probability outputs from the best transformer baseline (Wang and Manning 2012)(Fan et al. 2008).

Probability Calibration and Thresholding

Probability calibration was performed using Platt scaling on a validation subset(Platt 1999). Given the higher clinical cost associated with false negatives in the high-risk class, asymmetric decision thresholds were defined to prioritize recall for this class. Threshold selection was guided by precision-recall trade-offs and clinical safety considerations.(Saito and Rehmsmeier 2015)

Evaluation Metrics

As reported in Table 3, the transformer-based model (RoBERTuito) consistently outperformed the TF-IDF + Logistic Regression baseline across all evaluation metrics under 5-fold cross-validation. Specifically, RoBERTuito achieved higher accuracy (0.8323 vs. 0.7916), macro F1-score (0.7142 vs. 0.6893), and weighted F1-score (0.8315 vs. 0.8048). The improvement in macro F1-score indicates that the transformer model provides better balanced performance across all classes, including minority classes. Similarly, the higher weighted F1-score reflects improved overall predictive performance across the dataset.

At the class level, both models demonstrate strong performance in identifying low-risk cases ($F1 \approx 0.91-0.94$), reflecting the more explicit and homogeneous language associated with this category (see Fig. 2). However, the transformer model shows a clear advantage in the medium and high-risk classes, improving F1-scores from 0.62 to 0.65 and from 0.54 to 0.55, respectively. These improvements suggest that contextualized representations enable better capture of subtle and implicit patterns present in clinical narratives.

Despite these gains, performance in the high-risk class remains limited for both models, primarily due to lower recall, highlighting the inherent difficulty of detecting critical cases from unstructured text. Additionally, the higher standard deviation observed in minority classes indicates variability across folds, reflecting the challenges posed by data imbalance and limited sample size.

Table 3: Comparison of model performance across global metrics (Accuracy, Macro F1, Weighted F1) using 5-fold cross-validation.

Model	Accuracy	Macro F1	Weighted F1
TF-IDF + LR	0.7916	0.6893	0.8048
RoBERTuito	0.8323	0.7142	0.8315

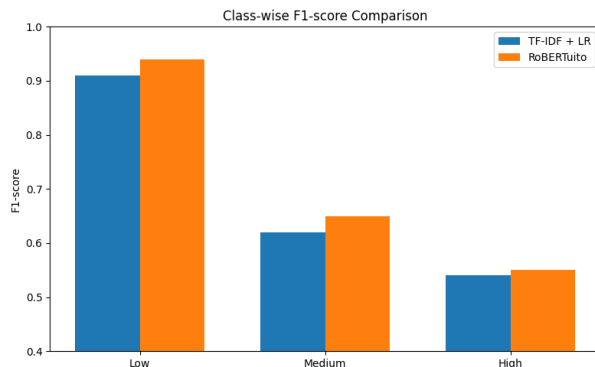


Figure 2: Comparison of F1-score by class. The transformer-based model (RoBERTuito) shows consistent improvements over the TF-IDF + Logistic Regression baseline, particularly in medium- and high-risk classes.

Results

The results describe the performance of the developed prototypes, using quantitative metrics, visualizations, and descriptive analyses.

Results of Therapeutic Plan Extraction

The therapeutic plan extraction component was evaluated on a manually annotated sample of over 100 psychiatric consultations using standard information extraction metrics (precision, recall, and F1-score). Five components were assessed, including pharmacological instructions, regimen changes, psychotherapy plans, follow-up actions, and medication refills.

The proposed pipeline achieved strong performance for most components, with F1-scores of 0.82 for medication-related information (Schema 1), 0.78 for psychotherapy plans (Schema 3), and 0.82 for follow-up actions (Schema 4). Medication refill detection (Schema 5) reached perfect performance, reflecting the explicit and standardized language used in clinical documentation.

In contrast, regimen change detection (Schema 2) showed low performance ($F1 = 0.15$), mainly due to the implicit and context-dependent nature of treatment modifications. Overall, the results indicate that the approach is effective for extracting explicit therapeutic information, while highlighting regimen change as a more complex and challenging task.

- **Schema 1: Medication + Dosage + Regimen**, representing a complete pharmacological instruction.
- **Schema 2: Regimen change (Medication + Regimen change)**, capturing adjustments to an existing treatment regimen.
- **Schema 3: Unified psychotherapy plan (consultation)**, describing the psychotherapy strategy documented during the visit.
- **Schema 4: Follow-up / control plan (consultation)**, indicating planned future visits or monitoring actions.

- **Schema 5: Medication refill (consultation)**, referring to explicit mentions of treatment renewal without modification.

As reported in Table 4, the proposed pipeline achieved strong performance for most components.

Table 4: Performance metrics for therapeutic plan extraction tasks.

Component	Precision	Recall	F1-score
Schema 1	0.79	0.85	0.82
Schema 2	0.09	0.39	0.15
Schema 3	0.78	0.78	0.78
Schema 4	0.82	0.82	0.82
Schema 5	1.00	1.00	1.00

Error Analysis Error analysis revealed recurring issues related to typing errors, non-standardized abbreviations, and implicit descriptions of treatment changes. These findings highlight the importance of clinical text quality as a key factor affecting the performance of free-text extraction systems.

Results of the Risk Signal Detection

One of the key findings of this study is that the transformer-based model (RoBERTuito) consistently outperforms the interpretable TF-IDF + Logistic Regression baseline across all evaluation metrics. While both approaches achieve strong performance, particularly in low-risk cases, the transformer model provides measurable improvements in overall accuracy and in macro-averaged F1-score, indicating better balanced performance across classes.

From a clinical perspective, the performance criterion was explicitly oriented toward prioritizing the detection of the negative (NEG) class, which represents higher-risk cases. Figure 3 provides a visualization of the trade-off between precision and recall for this class. The observed behavior supports prioritizing the detection of potentially relevant cases, even at the cost of accepting a moderate number of false positives.

In addition, the temporal evolution of the risk signal was explored for patients with multiple recorded consultations. For these cases, time series of the estimated risk probability were constructed, allowing for the observation of trends and changes over time. Figure 4 illustrates a representative example of this analysis, showing the variation of the risk signal across successive consultations.

Conclusion

This study shows that clinically relevant therapeutic information and indirect risk signals can be extracted from psychiatric narratives, even in low-resource settings. Interpretable models based on TF-IDF and logistic regression achieve competitive performance compared to transformer-based approaches, while offering greater transparency and stability, which are critical in clinical contexts. The use of calibrated probabilities and asymmetric thresholds further supports the prioritization of high-risk cases. Rather than

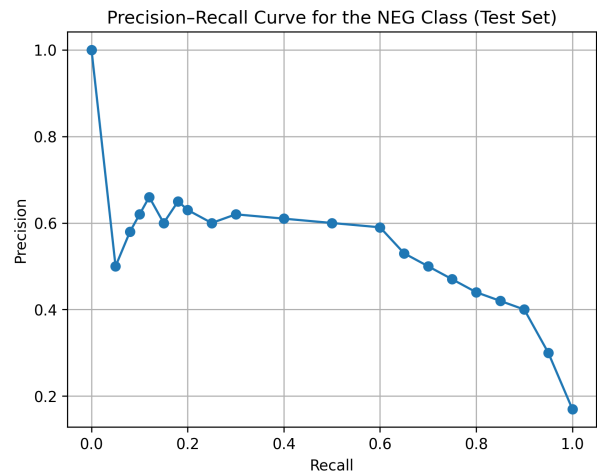


Figure 3: Accuracy-recall curve for the highest clinical risk class.

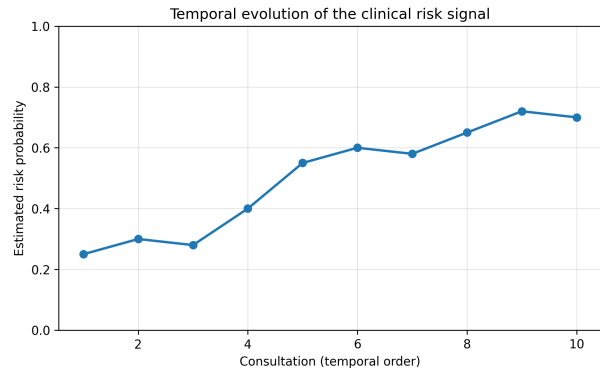


Figure 4: Temporal evolution of the estimated probability of clinical risk for a patient with multiple consultations.

modeling clinical risk directly, sentiment is used as a weak signal of underlying affective states. However, limitations remain, particularly in detecting regimen changes due to implicit and context-dependent language, as well as in the generalizability of results from a small, single-institution dataset. Overall, the findings highlight the importance of combining interpretability, clinical alignment, and robust evaluation when developing NLP tools for mental health.

Funding

This Project (No. PINV01-1224) is co-financed by the Paraguayan National Council of Science and Technology (CONACYT) with the support of the FEEL.

Ethical Considerations

This study was conducted in accordance with the ethical principles applicable to health data research, ensuring confidentiality, privacy, and the responsible use of clinical information. Data use was strictly limited to research purposes and had no impact on patient care.

References

- Camacho-Collados, J.; Pilehvar, M. T.; Collier, N.; and Navigli, R. 2017. Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, 15–26.
- Fan, R.-E.; Chang, K.-W.; Hsieh, C.-J.; Wang, X.-R.; and Lin, C.-J. 2008. Liblinear: A library for large linear classification. *the Journal of machine Learning research* 9:1871–1874.
- Le Glaz, A.; Haralambous, Y.; Kim-Dufor, D.-H.; Lenca, P.; Billot, R.; Ryan, T. C.; Marsh, J.; DeVyllder, J.; Walter, M.; Berrouguet, S.; and Lemey, C. 2021. Machine learning and natural language processing in mental health: Systematic review. *Journal of Medical Internet Research* 23(5):e15708.
- Navarro, G. 2001. A guided tour to approximate string matching. *ACM computing surveys (CSUR)* 33(1):31–88.
- Pérez, J. M.; Furman, D. A.; Alemany, L. A.; and Luque, F. M. 2022. Robertuito: a pre-trained language model for social media text in spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 7235–7243.
- Platt, J. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers* 61–74.
- Saito, T., and Rehmsmeier, M. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE* 10(3):e0118432.
- Wang, S. I., and Manning, C. D. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 90–94.
- World Health Organization. 2025. Over a billion people living with mental health conditions – services require urgent scale-up. <https://www.who.int/>. Accessed: 2026-02-01.