IJWR

Risk-Aware Suicide Detection in Social Media: A Domain-Guided Framework with Explainable LLMs

Farzaneh Lashgari*[a], Mehran Pourvahab[a] , António Sousa[a] , Anilson Monteiro [a], Sebastião Pais [a,b]
[a] NOVA-LINCS, University of Beira Interior, Covilha, Portugal; {f.lashgari, mehran.pourvahab, antonio.nave.sousa, anilson.monteiro, sebastiao}@ubi.pt
[b] Groupe de Recherche en Informatique, GREYC, University of Gaen Normandie, France;

ABSTRACT

Nowadays, the close connection between people's lives and social media has led to the emergence of their psychological and emotional states in social media posts. This type of digital footprint creates a rich and novel entry point for early detection of suicide risk. Accurate detection of suicidal ideation is a significant challenge due to the high false negative rate and sensitivity to subtle linguistic features. Current AI-based suicide detection systems are unable to detect linguistic subtleties. These approaches do not consider domain-specific indicators and ignore the dynamic interaction of language, behaviour, and mental health. Identifying lexical and syntactic markers can be a powerful diagnostic lens for diagnosing psychological distress. To address these issues, we propose a new domain-based framework that integrates the specialized frequent-rare suicide vocabulary (FR-SL) into the fine-tuning process of large language models (LLMs). This vocabulary-aware strategy draws the model's attention to common and rare suicide-related phrases and enhances the model's ability to detect subtle signs of distress. In addition to improving performance on various metrics, the proposed framework adds interpretability for understanding and trusting the models' decisions while creating transparency. It also enables the design of a structure that is generalizable to the linguistic and mental health domains. The proposed approach offers clear improvements over baseline methods, especially in terms of reducing false negatives and general interpretability through transparent attribution.

Keywords*— *Suicide Risk Detection, Large Language Models, Social Media Analysis, Mental Health Monitoring, Explainable AI

1. Introduction

The digital transformation of human interaction has reshaped the landscape of mental health expression. As individuals increasingly turn to social media for an emotional outlet, digital traces of psychological distress have emerged as potent indicators of suicidal ideation. This transition from private notes to public digital utterances signals a paradigm shift in how risk patterns manifest and how they must be understood and detected. Unlike clinical interviews or structured assessments, suicide-related expressions online are informal, context-sensitive, and often linguistically subtle, posing unique challenges for computational detection.

Despite significant advancements in clinical psychiatry and therapeutic tools [1], global suicide rates remain alarmingly high, particularly among youth populations [2]. Traditional prevention systems, including crisis hotlines and psychiatric interviews [3], frequently miss early distress indicators embedded within rapidly evolving digital discourse [4]. The failure of these systems is not due to negligence but rather to a mismatch between the linguistic richness of online communication and the tools available to interpret it [5]. Consequently, the detection of suicide risk has moved from being a purely psychological question to a computational and ethical challenge [6].

Artificial Intelligence (AI) [7], particularly Natural Language Processing (NLP) [8], has opened new avenues in computational mental health. The advent of transformer-based Large Language Models (LLMs)[9], such as GPT-NEO, Qwen2.5, and LLaMA3, has significantly advanced the ability to model language at scale [10]. These models extract meaning from context, capture dependencies

http://dx.doi.org/10.22133/ijwr.2025.525754.1288
Citation F. Lashgari, M. Pourvahab , A. Sousa, A. Monteiro, S. Pais, "Risk-Aware Suicide Detection in Social Media: A Domain-Guided Framework with Explainable LLMs", *International Journal of Web Research*, vol.8, no.3,pp.45-58, 2025, doi: http://dx.doi.org/10.22133/ijwr.2025.525754.1288.

**Coressponding Author*
Article History: Received: 24 April 2025 ; Revised: 9 June; Accepted: 26 June 2025 .

across long sequences, and adapt to varied writing styles [11]. However, their application in sensitive domains like suicide detection remains limited and often problematic [12]. These general-purpose models are rarely attuned to the subtle, low-frequency, domain-specific vocabulary associated with suicidal ideation [13,14]. Moreover, their black-box architecture raises critical concerns about interpretability and accountability in high-risk applications [15].

Three significant gaps underscore the limitations of existing approaches: 1- Generic LLMs often miss low-frequency linguistic cues that signal distress. For instance, phrases like “I am tired of trying” or “goodbye forever” may carry high clinical significance yet receive low attention due to sparse occurrence. 2- Existing models rarely explain why a particular post is flagged as high-risk, limiting trust in their predictions, especially when deployed in real-world mental health settings [16]. 3- Most models are trained on English-centric data with limited consideration for linguistic diversity or informal expressions that differ across communities and cultures.

To address these challenges, we proposed a domain-guided, explainable framework that augments LLMs with a custom-built Frequent-Rare Suicide Lexicon (FR-SL). This lexicon captures both common clinical expressions and informal, rare phrases indicative of suicidal ideation. By incorporating FR-SL into a two-phase fine-tuning pipeline, we sensitize LLMs to risk-relevant.

Linguistic patterns while preserving their general reasoning capabilities. Furthermore, to mitigate the opacity of LLM decisions, we integrate SHapley Additive explanations (SHAP) [17] for token-level interpretability, allowing human experts to trace predictions back to semantically meaningful cues.

We hypothesize that incorporating domain-specific lexical cues (FR-SL) into fine-tuning LLMs enhances suicide risk detection accuracy [18] and interpretability, particularly in detecting low-frequency and context-dependent suicidal expressions on social media.

This hypothesis is empirically evaluated through experiments on two Twitter-based datasets, employing three open-source LLMs under lexicon-augmented and baseline settings. Our goal is to improve classification metrics and build ethically transparent, linguistically sensitive, and resource-efficient models suitable for real-world deployment. Our contributions are:

- A domain-specific suicide lexicon encompassing high-frequency expressions and low-frequency informal indicators has been systematically constructed using a hybrid methodology that combines TF-IDF analysis with sentiment-aligned filtering. This lexicon enables models to become sensitized to both overt and subtle signs of suicidal ideation often overlooked in traditional settings.
- A two-stage fine-tuning strategy has been developed and applied to three representative open-source large language models (GPT-NEO, Qwen2.5, and LLaMA3). This strategy enables the controlled integration of domain-aware lexical cues, allowing for a direct comparison between standard LLM behaviour and lexicon-augmented performance under high-resource and low-resource conditions.
- The framework has been empirically validated on two benchmark Twitter-derived datasets, selected for their lexical diversity and expert-verified labels. Robustness, generalizability, and the impact of lexical integration were systematically evaluated through stratified cross-validation, quantitative metrics (accuracy, precision, recall, F1), and extensive error analysis.
- A SHAP-based explainability layer has been integrated to quantify token-level attribution and assess the interpretability of predictions. This layer enhances transparency and provides actionable insights into the semantic influence of lexicon terms, a critical feature for deployment in sensitive mental health contexts.
- This work advances the field by balancing algorithmic performance, explainability, and ethical readiness. It is positioned as a lightweight, resource-efficient, and culturally adaptable solution, suitable for deployment in real-world AI-driven mental health screening systems where interpretability and risk mitigation are paramount.

The rest of the paper is structured as follows: Section 2 reviews related literature and highlights existing gaps; Section 3 details the proposed methodology, including lexicon construction and model training; Section 4 presents results, error analysis, and explainability findings; and Section 5 concludes with ethical considerations and future directions.

2. Related Work

Suicide detection in social media has emerged as a critical and high-impact research domain at the intersection of artificial intelligence, linguistics, and public health. This surge is driven by the alarming increase in mental health challenges worldwide and

the ubiquity of user-generated content across platforms like Reddit, X (formerly Twitter), and Facebook. These platforms often serve as informal spaces where individuals disclose emotional distress, making them valuable yet complex resources for risk detection.

The earliest computational approaches to this problem predominantly employed classical machine learning (ML) techniques such as Support Vector Machines (SVM) [19], logistic regression, and decision trees. These models relied on handcrafted features, such as keyword lists and n-gram frequency patterns, which, although interpretable, could not be generalized across semantically diverse and context-dependent expressions. This led to high false-negative rates and poor robustness in detecting linguistic subtleties commonly observed in suicidal discourse [20]. Such models fail to capture emotional progression or implicit linguistic signals often embedded in figurative or idiomatic language.

A notable methodological shift is the adoption of deep learning architectures. Models such as Long Short-Term Memory (LSTM) networks [21] and Bidirectional Temporal Convolutional Networks (TCNs) [22] offered improved capacity to capture temporal and sequential dependencies in user posts. These systems were better suited to model emotional dynamics and recurring distress patterns over time. By integrating self-awareness mechanisms, the model's capability can be further enhanced by allowing flexible weighting of semantically critical tokens in large input sequences. However, these models still largely lacked key requirements such as transparency and domain sensitivity in high-risk applications such as suicide prevention.

Recent research has pivoted toward transformer-based models, particularly BERT and its domain-specific variants [23]. For instance, MentalBERT [24], trained on Reddit posts related to psychological discourse, showed enhanced contextual understanding of suicide-related text. However, despite improvements in accuracy, these models have been criticized for their black-box nature [25] and limited interpretability [26], which impedes clinical adoption and trustworthiness in real-world decision-making. The absence of transparent attribution mechanisms restricts the model's utility in ethically sensitive and high-stakes domains.

To address the need for adaptability in low-resource settings, Nguyen and Pham [27] proposed Mental-LLM, an instruction-tuned LLM capable of performing various mental health tasks. While promising in flexibility and low-data environments, this model lacked integration of domain-grounded lexical signals and did not offer interpretability tools, limiting its clinical applicability. Similarly, Quirch and El Ouazzani [28] introduced a hybrid deep learning model combining BiLSTM [29], CNN [30], and multiple word embeddings (Word2Vec, FastText, GloVe) [31], alongside transformer-based fine-tuning with BERT and GPT [32]. Their model achieved notably high classification accuracy (97.69%), showcasing the benefits of architecture fusion and linguistic diversity. However, this design remained opaque and resource-heavy.

Lasri et al. [33] further illustrated the effectiveness of large-scale LLMs by fine-tuning GPT-3 variants on the UMD Reddit Suicidality dataset and reaching an F1 Score of 92.3%. While these models deliver strong predictive performance, they function as resource-intensive black boxes [34], requiring large, annotated datasets and substantial computational infrastructure. These constraints limit their deployability in low-resource clinical or community settings [35], where interpretability, efficiency, and accessibility are paramount.

In contrast, our proposed framework addresses these limitations by emphasizing domain-guided lexical sensitivity, transparent prediction mechanisms, and lightweight design. We introduce a Frequent-Rare Suicide Lexicon (FR-SL) that encodes both clinically common and informally rare

linguistic indicators of suicidal ideation. By integrating this lexicon into a two-stage fine-tuning pipeline, we steer the attention of LLMs, specifically GPT-Neo, Qwen2.5, and LLaMA3, toward semantically meaningful risk cues. This approach enhances detection accuracy and facilitates token-level interpretability through SHapley Additive explanations (SHAP), allowing practitioners to trace model predictions back to identifiable linguistic evidence.

Unlike prior systems that prioritize performance over transparency or require extensive deployment resources, our method provides a balanced, culturally adaptable, and clinically aware solution well-suited for real-world integration in mental health monitoring tools where trust, clarity, and linguistic nuance are essential. Table 1 shows the comparison of related works on suicide detection.

3. Methodology

This section outlines our domain-specific methodology for suicide risk detection on social media using lexicon-enhanced LLMs. The proposed framework addresses two critical challenges in this domain: (1) the subtle and context-dependent nature of suicide-related expressions, and (2) the need for interpretable model behaviour in high-risk applications. A complete schematic of the end-to-end workflow from raw tweet collection to explainable model prediction is depicted in Figure 1, which illustrates the interaction between linguistic filtering, FR-lexicon integration, and transformer-based classification.

3.1. Dataset

To support model training and evaluation, we utilized two Twitter-derived datasets selected for their high relevance, lexical diversity, and validated annotation protocols targeting suicidal ideation. The rationale for selecting both corpora lies in their complementary characteristics: one emphasizes the breadth and variety of suicide-related expressions, while the other ensures expert-labelled clinical validity for benchmarking risk classification models.

Table 1. Comparison of Suicide Detection Models in social media

Model	***Dataset***	***Key Features***	***Reference***
Mental-LLM	Reddit QA (Mental Health)	Instruction-tuned LLM; fast adaptation; lacks interpretability	[27]
MentalBERT	Reddit posts	Domain-tuned BERT; rich contextual embeddings; lacks explainability	[23]
C-BiLSTM + GPT	Reddit (Mental Health)	BiLSTM + CNN + Word2Vec/FastText/GloVe + GPT; accuracy up to 97.69%	[28]
GPT-3 Fine-tuned	UMD Reddit Suicidality	Suicide risk-level detection; F1-score 92.3%; interpretable outputs	[33]
Domain-aware models (GPT-Neo, LLaMA3, Qwen2.5)	Various social media datasets	Domain-guided; explainable; resource-efficient; robust in low-resource settings	Our Work

Dataset1 (Explicit Suicidal Intent Set):

This dataset consists of 9,119 public tweets containing explicit or implicit suicidal content, collected via targeted keyword queries such as want to end it" or "no reason to live". The dataset was originally curated to encompass overt expressions (e.g., direct statements of intent) and covert signals (e.g., metaphorical or ambiguous language). For our study, we extracted a refined lexical subset to support the design and weighting of the FR Suicide Lexicon [36].

Dataset2 (Twitter Suicide Risk Corpus):

This dataset includes 3,200 English tweets manually annotated by a mental health expert and two computational linguists. Tweets were classified based on their alignment with established psychological risk indicators, resulting in a binary categorization: suicide risk vs. non-risk. The annotation framework aligns with the risk categorization guidelines used in contemporary benchmarking tasks for mental health NLP systems [37].

Together, these datasets provide a robust foundation for training, lexicon extraction, and performance evaluation under both high- and low-resource lexical conditions.

3.2. Preprocessing and Intent Realignment

The raw tweets were subjected to a standardized preprocessing [38] pipeline to normalize and sanitize the text for subsequent model training. All textual content was lowercased to ensure consistency across tokens. Emojis, hyperlinks, and extraneous punctuation marks were removed to reduce noise and eliminate artifacts irrelevant to semantic interpretation. Repetitive word forms, such as elongated tokens and repeated phrases, were deduplicated to minimize stylistic redundancy. In

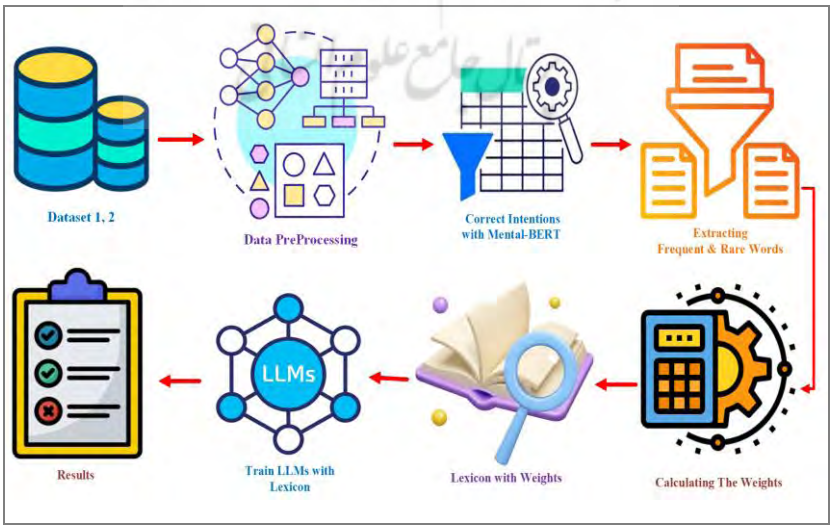


Figure. 1. Overall framework of our work

contrast, non-informative tokens, including common stopwords, were filtered out to sharpen the focus on semantically rich content. Beyond these surface-level transformations, an additional layer of refinement was applied through an intent realignment process. Given the prevalence of sarcasm, metaphor, and indirect language in expressions of suicidal ideation on social media, some instances in the dataset were likely to be mislabelled or ambiguously annotated. To address this, an auxiliary classifier based on MentalBERT, a BERT variant pre-trained on mental health-related corpora, was fine-tuned on a curated subset of high-confidence suicidal posts.

This domain-aware model was then employed to reassess borderline or ambiguous samples, particularly those exhibiting stylistic markers such as irony, figurative language, or emotional contradiction. By leveraging the contextual sensitivity of MentalBERT, the realignment process significantly improved the semantic reliability of the annotations and reduced the likelihood of false negatives in downstream training.

Moreover, to alleviate class imbalance and enhance the representation of the minority class (i.e., suicidal posts), additional high-risk tweets were selectively incorporated into the dataset. This ensured that the resulting training corpus maintained lexical diversity and representational adequacy for robust suicide risk detection.

3.3. Frequent-Rare Suicide Lexicon Construction

A core innovation of this study lies in the development of the FR-SL, a domain-specific vocabulary crafted to improve the model's sensitivity to linguistic cues indicative of suicide risk. Unlike general-purpose lexicons, FR-SL was designed to capture common and infrequent phrases that reflect mental health distress, particularly as they manifest in social media discourse.

The lexicon construction process was based on a frequency-based thresholding method, followed by sentiment-informed filtering. First, we used Term Frequency-Inverse Document Frequency (TF-IDF) [39] analysis to extract candidate terms from both datasets. Words that appeared more than 200 times across the corpora were labelled as frequent, while those occurring between 10 and 15 times were considered rare. To assess the salience of these terms, we calculated TF-IDF scores for each group to highlight their contextual significance within the suicide-related subsets.

Next, we intersected the frequent and rare term lists with entries from the NRC Emotion Lexicon [40], a widely used sentiment dictionary that assigns affective labels (e.g., sadness, fear, anger) to words. Only terms present in our filtered set and the NRC lexicon were retained for inclusion in FR-SL. Finally, we applied a weighting adjustment based on Equation (1).

$$w_i = \lambda_i \cdot \text{TF-IDF}(i) + \sum_{s \in S_i} \delta_s \qquad (1)$$

Where:
- w_i denotes the final weight assigned to term *i*;
- **TF-IDF(*i*)** is the term frequency-inverse document frequency score of term *i*;
- S_i represents the set of sources confirming the semantic or emotional relevance of term *i*;
- λ_i is a proportional coefficient reflecting the number of overlapping sources in which term *i* appears;
- δ_s indicates the degree of alignment between source s and domain-specific or sentiment-based relevance.

Unlike other approaches, our lexicon does not rely on predefined clinical phrases or external sentiment scores. Instead, it leverages corpus-internal frequency cues and domain-aligned affective overlaps to identify semantically important expressions ranging from overt signals (e.g., "kill myself") to more subtle indicators (e.g., "I cannot anymore", "goodbye forever"). This strategy allowed us to construct a lightweight, interpretable, and contextually grounded lexicon suitable for integration into LLM fine-tuning. This weighting ensured that semantically critical terms such as *"goodbye forever"* or *"can't go on"* received higher emphasis, while non-specific words like *"bad day"* were demoted unless corroborated.

3.4. Model Architecture and Training Strategy

To evaluate the impact of lexicon-guided fine-tuning on suicide risk detection, we employed three representative open-source large language models: GPT-NEO, Qwen2.5, and LLaMA3. These models were selected based on their architectural diversity and tokenizer strategies, which reflect different capacities for handling informal, noisy, and compressed textual data typical of social media platforms. Specifically, GPT-NEO utilizes a byte-level BPE tokenizer, Qwen2.5 is equipped with a standard BPE tokenizer, and LLaMA3 adopts a hybrid SentencePiece-based Unigram+BPE tokenizer. All datasets were preprocessed and tokenized using each model's native tokenizer to preserve semantic fidelity and structural consistency.

The models were fine-tuned using a two-stage training strategy to enable controlled integration of domain-specific lexical cues. In the baseline stage, they were trained on tokenized input sequences using a standard binary classification objective without lexical supervision. This stage served as a benchmark for evaluating each model's inherent

ability to learn from generic training signals in this study.

In the lexicon-augmented phase, the models already trained in the baseline setting were further fine-tuned by incorporating the curated FR-SL. Rather than injecting lexical guidance from the outset, we opted for delayed integration to prevent premature overfitting and allow the model to develop general linguistic representations first. During this phase, tokens that matched entries in the FR-SL were encoded with auxiliary semantic tags during preprocessing. Their corresponding embeddings were lightly re-initialized to increase attention weights, effectively nudging the model toward clinically meaningful risk indicators without disrupting previously learned context patterns. This strategy allowed the models to retain their general reasoning capabilities while becoming more sensitive to subtle suicide-related linguistic signals.

To enhance the results' robustness and generalizability, we applied 5-fold stratified cross-validation in all experiments and also enabled early stopping based on loss of validation. This helped to reduce overfitting, especially given the lexicon weight class imbalance and lexical scarcity inherent in real-world social media datasets. This two-stage training approach allowed for a fair and controlled comparison between each model architecture's standard and lexical-enhanced learning pipelines.

3.5. Model Selection Justification

The selection of GPT-NEO, Qwen2.5, and LLaMA3 was guided by a combination of architectural diversity, tokenizer characteristics, and open-access availability. GPT-NEO was chosen due to its byte-level BPE tokenizer, which excels in handling noisy, informal, and stylistically diverse text typical of social media platforms. Its conservative attention mechanism also favours high-confidence predictions, making it suitable for minimizing false positives. Qwen2.5 represents a balance between compact encoding and adaptability, offering strong performance with fewer computational resources. LLaMA3, with its hybrid SentencePiece tokenizer (Unigram + BPE), was included to assess the ability of flexible token segmentation in capturing rare or fragmented linguistic patterns.

These models were prioritized over other mainstream LLMs (e.g., BERT, RoBERTa, ChatGPT) due to their open-source availability, lightweight architectures, and fine-tuning flexibility. Unlike ChatGPT and similar closed models, they allow direct intervention in token embeddings and lexicon integration. Compared to BERT-based models, our selected LLMs better support long-context dependencies and cross-token semantic alignment, which are crucial for capturing subtle suicidal cues. This diversity ensures a more comprehensive evaluation of how lexicon-guided fine-tuning performs across different transformer configurations.

3.6. Implementation Details and Computational Resources

All experiments were conducted using PyTorch and the HuggingFace Transformers library on a single NVIDIA RTX 3090 GPU (24GB VRAM). Each model GPT-NEO, Qwen2.5, and LLaMA3 was fine-tuned under both baseline and lexicon-augmented conditions across two datasets using 5-fold stratified cross-validation. Training was performed using a binary cross-entropy loss, with early stopping based on validation loss. The learning process employed a batch size of 8, learning rate of 2e-5, and up to 10 epochs per fold, though early stopping typically halted training earlier. SHAP-based token-level explainability was performed after training on each model. The entire pipeline including data preprocessing, lexicon integration, fine-tuning, evaluation, and explainability accumulated to approximately 450 GPU hours in total. No architectural modifications were applied to the models; however, tokenizer-specific preprocessing and lightweight embedding reinitialization were applied for lexicon-aligned tokens. All experiments were version-controlled and fully reproducible, ensuring transparency and traceability.

4. Results and Discussion

To evaluate the impact of lexicon-guided fine-tuning, we conducted a series of experiments using GPT-NEO, Qwen2.5, and LLaMA3 across two distinct datasets. The results demonstrate that incorporating the FR-SL significantly improves suicide risk detection performance, especially regarding recall and F1-score. This section discusses the quantitative findings, model behaviour under lexical constraints, and insights derived from error analysis and SHAP-based explainability.

4.1. Performance on Dataset 1: Lexical Diversity and Signal Density

In our previous work [18], we established a foundational benchmark for suicide risk detection using large language models, reporting promising but model-specific performance trends across diverse architectures. Building upon those findings, we extended the experimental setting on Dataset 1 (n = 9,119), which features a broader spectrum of suicidal expressions, to examine the generalizability of lexicon-guided augmentation using the FR-SL framework.

The integration of the FR-SL lexicon yielded consistent performance improvements across all three models, as presented in Table 2 and visualized

in Figure 2. Notably, Qwen2.5 achieved the highest F1-score of 92.68% following augmentation, up from 46.08% in the base setting. It also showed substantial improvements in accuracy (92.93% vs. 30.42%), precision (94.21% vs. 37.18%), and recall (91.20% vs. 60.60%), highlighting its strong adaptability to rare suicidal cues—likely aided by its balanced BPE tokenization.

GPT-NEO exhibited the most balanced overall performance in the FR-SL setting, achieving an F1-score of 93.40%, up significantly from 44.87% in the base condition. Accuracy improved from 63.10% to 93.52%, precision from 84.07% to 93.40%, and recall from 30.60% to 93.40%. Its conservative byte-level tokenizer appears to support high-confidence classification, reducing false positives while preserving sensitivity.

LLaMA3, while starting from a lower baseline, also benefited from lexicon-guided prompting. Its F1-score rose from 33.12% to 91.02%, with corresponding increases in accuracy (48.87% to 91.17%), precision (46.24% to 90.84%), and recall (25.80% to 91.20%). Nevertheless, its improvements were slightly more constrained compared to the other models, possibly due to limitations introduced by its hybrid tokenizer, which may fragment rare lexical patterns.

Overall, these results validate and extend the observations made in SENTINEL-LLM, confirming that lexicon-informed prompting (via FR-SL) is an effective enhancement strategy across diverse LLM architectures. Furthermore, the comparative trends emphasize the role of tokenizer design and model alignment in downstream suicide risk detection tasks.

Table 2. Performance Comparison of Base vs. SENTINEL LLM Models on Dataset 1[18]

Model	*Setting*	*Acc.*	*Prec.*	*Recall*	*F1*
GPT-NEO	Base	63.10	84.07	30.60	44.87
	FR-SL	**93.52**	93.40	**93.40**	**93.40**
Qwen2.5	Base	30.42	37.18	60.60	46.08
	FR-SL	92.93	**94.21**	91.20	92.68
LLaMA3	Base	48.87	46.24	25.80	33.12
	FR-SL	91.17	90.84	91.20	91.02

4.2. Performance on Dataset 2: Low-Resource Settings

In Dataset 2 (*n = 3,200*), which featured fewer lexical variants but higher annotation precision, the overall performance of baseline models declined, largely due to data sparsity and narrower distribution of suicidal cues. Nonetheless, the integration of the FR-SL lexicon consistently enhanced all four-evaluation metrics across models, as shown in Table 3 and Figure 3.

GPT-NEO retained the highest precision (91.53%) and demonstrated strong gains in recall (55.08% → 91.53%) and F1-score (54.85% → 85.04%) following FR-SL augmentation. These results highlight its robustness and conservative generalization behavior, even under lexical compression.

Qwen2.5, which had struggled in the baseline setting, particularly in recall (11.02%) and F1-score (17.69%), exhibited the highest post-augmentation F1-score (91.49%), driven by sharp improvements in both precision (93.48%) and recall (89.58%). However, its extreme sensitivity to lexicon absence suggests reliance on semantic guidance for low-resource data contexts.

LLaMA3 achieved more stable performance across settings, improving from a baseline F1-score of 66.67% to 82.35% with FR-SL. Although less extreme than Qwen2.5, LLaMA3's recall variation (67.46% → 83.05%) still indicates potential challenges with idiomatic or compact phrasing, possibly linked to hybrid token segmentation.

These outcomes reinforce the value of FR-SL augmentation, particularly in data-scarce scenarios, and illustrate varying degrees of lexicon dependency and tokenizer sensitivity among LLM architectures.

4.3. Assessing Vocabulary Impact on Model Performance

By replicating the experiments on Dataset 2, which contains roughly one-third the size of Dataset

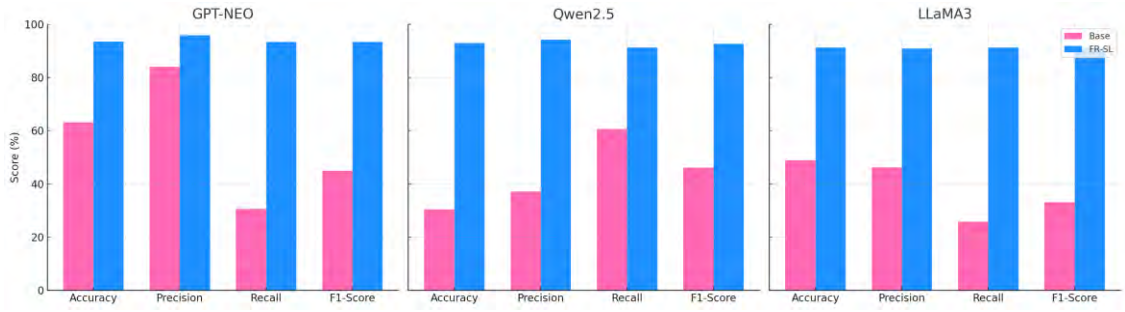


Figure. 2. Performance comparison of 3 models on dataset 1.

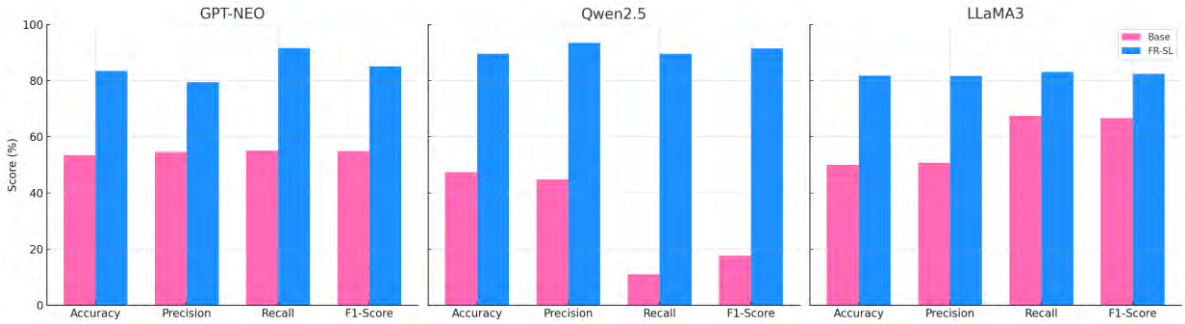


Figure. 3. Performance comparison of 3 models on dataset 2.

Table 3. Performance Comparison of Base vs. FR-SL fine-tune Models on Dataset 2

Model	*Setting*	*Acc.*	*Prec.*	*Recall*	*F1*
GPT-NEO	Base FR-SL	53.48 83.48	54.62 79.41	55.08 **91.53**	54.85 85.04
Qwen2.5	Base FR-SL	47.39 **89.57**	44.83 **93.48**	11.02 89.58	17.69 **91.49**
LLaMA3	Base FR-SL	50.00 81.7۴	50.66 81.67	67.46 83.05	66.67 82.35

1 (n = 3,200), we created an opportunity to empirically examine the relationship between lexical diversity and the effectiveness of lexicon-guided augmentation. The comparative results clearly indicate that as the variety of expressive language, and consequently the coverage of the FR-SL lexicon, decreases, the impact of lexicon-enhanced fine-tuning on model performance also diminishes. This trend is particularly evident for models like Qwen2.5, which demonstrated dramatic performance gains in Dataset 1 but showed heightened sensitivity to lexical sparsity in Dataset 2. These findings underscore that the effectiveness of lexical augmentation is not only dependent on model architecture and tokenizer design but is also strongly conditioned by the lexical richness of the training data and its overlap with the lexicon. In essence, with reduced lexical variety, the model's ability to leverage semantic guidance from the lexicon becomes constrained.

From a methodological perspective, this comparative experiment offers a significant contribution by empirically demonstrating that lexicon-based fine-tuning functions as a domain-sensitive augmentation strategy. Its effectiveness is modulated by the lexical and semantic density of the target dataset, indicating that the extent of vocabulary overlap between the lexicon and dataset can serve as an indirect proxy for estimating the semantic informativeness of the data. This insight not only informs future lexicon construction but also highlights the potential of lexicon-driven augmentation as a diagnostic tool for evaluating the representational depth of low-resource corpora.

4.4. Linguistic and Architectural Interpretation

The observed differences between models are not merely statistical but deeply rooted in tokenizer design, model architecture, and vocabulary alignment strategies:

- With its Byte-level BPE tokenizer, GPT-NEO excels in managing noisy, informal, or non-standard inputs key features of social media data. Its architecture, being more deterministic in handling tokens, likely supports higher precision but at the cost of missing nuanced low-frequency patterns unless explicitly trained.
- Qwen2.5, built on a BPE tokenizer, strikes a middle ground: it compresses frequent sequences efficiently while maintaining generalizability. Its superior F1 score, especially after FR-SL integration, indicates high adaptability to vocabulary-driven fine-tuning. This property makes it especially suitable for real-world social media screening tasks.
- LLaMA3, incorporating a Unigram + BPE approach, theoretically offers flexibility to represent rare and common words. However, its performance suggests that such flexibility may need more fine-tuned calibration when applied to short, noisy, and semantically dense texts like those on social platforms.

4.5. Relevance of Domain-Specific Vocabulary

The performance gains across both datasets validate the effectiveness of domain-specific vocabulary integration. Unlike standard fine-tuning, our approach emphasizes lexicon-guided alignment, targeting both frequent clinical terms (e.g., "suicidal ideation") and informal social cues (e.g., "done with life", "I can't anymore") that appear outside traditional clinical narratives.

Such alignment is critical because social media language fundamentally differs from clinical interviews: it often involves abbreviations, metaphors, sarcasm, slang, and context-dependent expressions not present in structured medical settings. The FR-SL bridges this gap by mapping high-frequency clinical phrases to their low-frequency, informal counterparts, improving model coverage and interpretability.

4.6. Error Analysis

To deepen our understanding of model behaviour beyond aggregate metrics such as accuracy and F1-score, we conducted a comparative error analysis focusing on false positives (FP) and false negatives (FN) across both datasets. Table 4 presents a unified overview of classification errors for each model under both baseline and FR-SL-augmented settings. Reduction in Total Errors with Lexicon Augmentation: All three models substantially reduced total errors (FP + FN) when augmented with the FR lexicon. For example, GPT-NEO's total errors dropped from 4895 to 947 in Dataset 1, and from 1814 to 644 in Dataset 2.

Shift in Error Trade-offs

Lexicon augmentation reduced total errors and reshaped the balance between FP and FN. In the baseline condition, most models specially Qwen2.5 and LLaMA3 struggled with high FN due to implicit suicide cues. After FR-SL integration, FN rates dropped significantly (e.g., LLaMA3: 4394 → 763 in Dataset 1), even though FP slightly increased in some cases.

Model-Specific Behavior

GPT-NEO maintained a relatively balanced error profile with consistent improvements across both datasets. Qwen2.5, initially prone to high FN in Dataset 1, showed the most dramatic reduction, especially in Dataset 2 (total errors: 333). LLaMA3 exhibited unusually high FN in its baseline form (4394) but significantly recovered after lexicon integration.

Model-Specific Behavior

GPT-NEO maintained a relatively balanced error profile with consistent improvements across both datasets. Qwen2.5, initially prone to high FN in Dataset 1, showed the most dramatic reduction, especially in Dataset 2 (total errors: 333). LLaMA3 exhibited unusually high FN in its baseline form (4394) but significantly recovered after lexicon integration.

Dataset Sensitivity and Generalization

Despite the smaller scale of Dataset 2, relative improvements remained consistent, highlighting that lexicon-guided learning generalizes well under both high-resource and low-resource conditions. Notably, lexicon augmentation also reduced FP, where many false positives in baseline runs involved words like "die" or "kill" in non-suicidal contexts (e.g., movie quotes). The FR-SL helped models better disambiguate these by guiding attention to semantically relevant tokens.

4.7. Explainability and Model Interpretation

To achieve interpretability and better understand the decision-making processes of the language models when utilizing the FR-dictionary vocabulary, we employed SHAP analysis. SHAP values quantify the contribution of each token toward the final prediction, providing insights into the relative influence of domain-specific vocabulary. Figures 4 - 6 illustrate token attribution distributions for GPT-NEO, Qwen2.5, and LLaMA3.

Our SHAP analyses revealed meaningful differences in how each model utilizes FR-dictionary terms. The average contribution of FR tokens across models was around 9% (mean) and 5% (median), with some samples exceeding 60%. In contrast, ~24% of GPT-NEO predictions showed no FR-token contribution, suggesting inference based on broader context.

Table 4. Error Counts (FP, FN, Total) Across Datasets and Settings

Model	*Setting*	*Dataset1*			*Dataset2*		
		FP	*FN*	*Total*	*FP*	*FN*	*Total*
GPT-NEO	Baseline	447	4448	4895	915	898	1814
	With FR-SL	184	763	947	475	169	644
Qwen2.5	Baseline	3157	2395	5552	271	1780	2051
	With FR-SL	263	606	869	125	208	333
LLaMA3	Baseline	1053	4394	5447	1898	51	1949
	With FR-SL	398	633	1031	672	340	1012

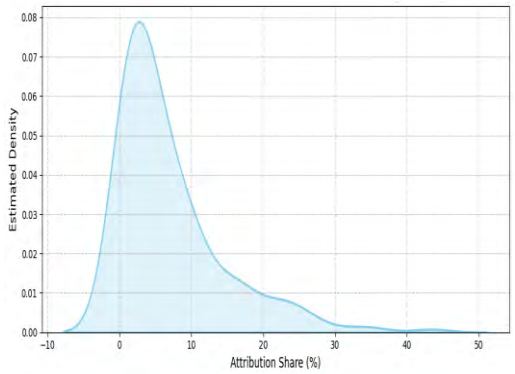


Figure. 4. SHAP token attribution analysis for GPT-NEO

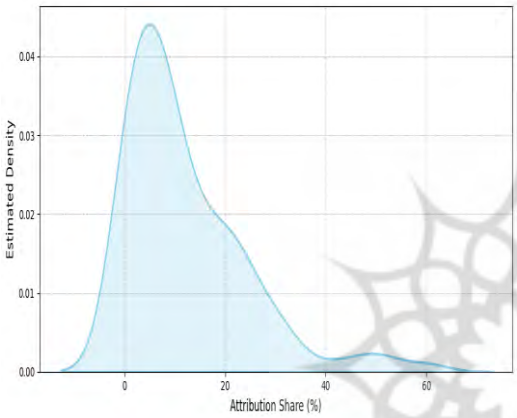


Figure. 5. SHAP token attribution analysis for Qwen2.5

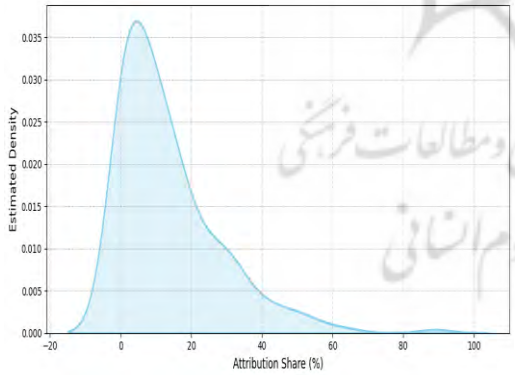


Figure. 6. SHAP token attribution analysis for LLaMA3

Model-specific patterns were also evident: GPT-NEO showed cautious yet focused attribution spikes; Qwen2.5 and LLaMA3 demonstrated smoother, more distributed token attribution, reflecting a balance between explicit vocabulary cues and contextual understanding. These findings underscore the interpretability gains enabled by lexicon integration. Predictions with zero attribution to FR tokens highlight opportunities for expanding the lexicon or improving contextual learning strategies. Overall, the SHAP-based interpretability framework confirms the effectiveness of vocabulary-guided learning and enhances transparency in model decisions an essential feature for clinical and real-world deployment in sensitive domains such as suicide prevention.

4.8. Limitations

While our framework presents notable advancements in detecting suicidal ideation, several limitations remain that open promising avenues for future work, many of which can be addressed within the current architectural scope.

First, our study relies exclusively on publicly available social media data. Although this ensures transparency and replicability, it excludes posts from private or semi-private communities, where individuals may disclose more intense psychological distress. Future extensions could incorporate privacy-preserving collaboration mechanisms with platform providers or employ federated learning to analyze sensitive content without compromising user confidentiality.

Second, our framework is currently restricted to textual content. However, suicidal ideation is often manifested through multimodal signals, such as vocal tone, facial expression, and visual symbols. Given our model's modular structure, it can be extended to integrate cross-modal embeddings and pre-trained visual/audio encoders. For instance, audio-to-text transcription can be combined with emotional prosody analysis, or image-based sentiment models can be fused using late-stage attention mechanisms. This would enable a more holistic understanding of distress beyond language.

Third, all experiments were conducted in English. This inherently limits the generalizability of our findings across linguistically diverse populations. To address this, the existing framework can incorporate multilingual pre-trained LLMs (e.g., XLM-R or mBERT) and align the FR-SL lexicon with language-specific suicide-related expressions via embedding projection or translation-guided lexical alignment. In future work, cross-cultural corpora annotated by multilingual experts could be used to refine and validate such extensions.

Fourth, the construction of the FR-SL lexicon involved semi-automated techniques, blending frequency thresholds with manual filtering. While this yielded high interpretability, it limits scalability across domains or languages. The current pipeline can be expanded using automated vocabulary expansion techniques, such as embedding similarity, self-supervised clustering, and dynamic querying of LLMs under emotion-conditioned prompts. These enhancements would enable continuous updates and cultural adaptability of the lexicon.

Finally, although our system includes explainability mechanisms (via SHAP), further improvements could involve interactive explanation interfaces for clinical users or uncertainty quantification modules to flag ambiguous cases. These additions would increase the trustworthiness and accountability of deployment in sensitive contexts.

In summary, while some limitations persist, they do not reflect fundamental flaws in the proposed architecture. Rather, they highlight directions for enriching the system's multimodal, multilingual, and adaptive capabilities within the same explainable and domain-guided design philosophy.

4.9. Ethical Considerations

All datasets used in this study were publicly available, anonymized, and selected in compliance with ethical standards for research involving human-related content. No personally identifiable information (PII) was accessed or utilized, and no direct interaction with individuals occurred at any study stage. The proposed system is intended solely as a tool to support faster and more accurate identification to assist clinicians, administrators, and mental health professionals, but is in no way guaranteed to replace human judgment.

Furthermore, the integration of explainability layers such as SHAP ensures that the model's predictions remain transparent, debatable, and auditable by human experts. This transparency is particularly critical in high-risk applications such as suicide prevention, where false positive or negative results have serious ethical and psychological consequences. Our design takes into account the principles of responsible AI and promotes safety, fairness, and accountability in the deployment of computational systems in vulnerable domains.

5. Conclusions

This study proposed a domain-aware, lexicon-guided framework (FR-SL) that enhances suicide risk detection on social media by integrating semantically rich vocabulary into large-scale language models (GPT-NEO, Qwen2.5, and LLaMA3). Our experimental results on two benchmark datasets demonstrate significant improvements in recall and F1 scores, which align with the effectiveness of fine-tuning LLMs with domain-specific lexical signals. These improvements were particularly prominent in identifying rare and context-specific indicators of suicidality that conventional models often miss.

Beyond the quantitative results, the incorporation of SHAP-based explainability has enabled us to relate model predictions to specific lexical stimuli, and to have transparency around the interpretation and operationalization of the model particularly in sensitive applications like mental health monitoring systems. Combining an explicit rationale that aligns model behavior with clinical reasoning pathways allows us to ground the raw algorithmic power to its real-world applications, and opens up a number of opportunities for future work. First, expanding the coverage of the FR-SL lexicon via automated discovery methods such as embedding-based similarity, dynamic querying of LLMs, and clustering co-occurrence statistics could improve adaptability to emerging linguistic trends. Second, multilingual and culturally contextual extensions are essential for global deployment, especially in regions with high linguistic variation in the expression of psychological distress. Thirdly, by tapping into multimodal data like emotional cues from voices, the feelings conveyed in images, and behavioral metadata such as how often someone posts or interacts we can really boost the accuracy and strength of risk detection. This approach blends sensitivity to language, clarity in understanding, and efficiency in models, creating a solid base for responsibly integrating AI into the field of computational mental health. Looking ahead, advancements in this area hold tremendous promise for broadening the reach and dependability of suicide prevention initiatives across various cultural and technological contexts.

Declarations

Funding

This research did not receive any grant from funding agencies in the public, commercial, or non-profit sectors.

Authors' contributions

All authors contributed equally to the study conception, design, data analysis, interpretation of results, and manuscript preparation. All authors read and approved the final manuscript.

Conflict of interest

The authors declare that no conflicts of interest exist.

Acknowledgements (optional)

This work is supported by UID/04516/NOVA Laboratory for Computer Science and Informatics (NOVA LINCS) with the financial support of FCT.IP

References

[1] C. Salvador, V. Felizardo, H. Zacarias, L. Souza-Pereira, M. Pourvahab, N. Pombo, and N. M. Garcia, "Epileptic seizure prediction using EEG peripheral channels," in *2023 IEEE 7th Portuguese Meeting on Bioengineering (ENBENG)*, Porto, Portugal, 2023, pp. 60–63. https://doi.org/10.1109/ENBENG58165.2023.10175347

[2] S. Gabriel, I. Puri, X. Xu, M. Malgaroli, and M. Ghassemi, "Can AI relate: Testing large language model response for mental health support," in Proc. Findings of the Association for Computational Linguistics: EMNLP

2024, 2024, pp. 2206–2221. https://doi.org/10.18653/v1/2024.findings-emnlp.120

[3] D. Shin, H. Kim, S. Lee, Y. Cho, and W. Jung, "Using large language models to detect depression from user-generated diary text data as a novel approach in digital mental health screening: Instrument validation study," *J. Med. Internet Res.*, vol. 26, p. e54617, Sep. 2024. https://doi.org/10.2196/54617

[4] A. Chowdhury, K. Mukhopadhyay, O. Ganguly, and A. Maiti, "Enhancing mental health support through AI-enabled chatbots: Integrations, impacts, and future directions," in *14th Inter-University Engineering, Science & Technology Academic Meet*, 2024. https://www.researchgate.net/publication/381772916

[5] T. Galijašević, M. Škarić, E. Podolski, F. Mustač, M. Matovinović, and D. Marčinko, "New breakthroughs in AI chatbots and their potential in mental health services," in *2024 9th Int. Conf. Smart and Sustainable Technologies (SpliTech)*, Bol and Split, Croatia, 2024, pp. 01–04. https://doi.org/10.23919/SpliTech61897.2024.10612516

[6] T. H. Sakib, M. Ishak, F. F. Jhumu, and M. A. Ali, "Analysis of suicidal tweets from Twitter using ensemble machine learning methods," in *Proc. ACM Int. Conf. Mach. Learn. (ACMI)*, Rajshahi, Bangladesh, IEEE, Jul. 2021, 1-7. https://doi.org/10.1109/ACMI53878.2021.9528252

[7] F. Muetunda, S. Pais, S. Sabry, G. Dias, N. Pombo, and J. Cordeiro, "Improving mental disorder predictions using feature-based machine learning techniques," in *Proc. 2023 23rd IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Shanghai, China, Dec. 2023, pp. 1279–1288. https://doi.org/10.1109/ICDMW60847.2023.00165

[8] M. Pourvahab, S. J. Mousavirad, V. Felizardo, N. Pombo, H. Zacarias, H. Mohammadigheymasi, S. Pais, S. N. Jafari, and N. M. Garcia, "A cluster-based opposition differential evolution algorithm boosted by a local search for ECG signal classification," *J. Comput. Sci.*, vol. 86, p. 102541, 2025. https://doi.org/10.1016/j.jocs.2025.102541

[9] Z. Liu, Y. Bao, S. Zeng, R. Qian, M. Deng, A. Gu, J. Li, W. Wang, W. Cai, W. Li, H. Wang, D. Xu, and G. N. Lin, "Large language models in psychiatry: Current applications, limitations, and future scope," *Big Data Min. Anal.*, vol. 7, no. 4, pp. 1148–1168, 2024. https://doi.org/10.26599/BDMA.2024.9020046

[10] H. R. Lawrence, R. A. Schneider, S. B. Rubin, M. J. Matarić, D. J. McDuff, and M. J. Bell, "The opportunities and risks of large language models in mental health," *JMIR Ment. Health*, vol. 11, p. e59479, Jul. 29, 2024. https://doi.org/10.2196/59479

[11] U. Sirisha and B. Chandana, "Aspect based sentiment and emotion analysis with ROBERTa, LSTM," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, Jan. 2022. https://doi.org/10.14569/IJACSA.2022.0131189

[12] A. Kermani, V. Perez-Rosas, and V. Metsis, "A systematic evaluation of LLM strategies for mental health text analysis: Fine-tuning vs. prompt engineering vs. RAG," in *Proc. 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, San Marcos, TX, USA: Association for Computational Linguistics, May 2025, pp. 172–180. https://aclanthology.org/2025.clpsych-1.14

[13] H. Zacarias *et al.*, "Gender classification using nonstandard ECG signals – A conceptual framework of implementation," in *IoT Technologies for HealthCare*, Cham, Switzerland: Springer Nature Switzerland, 2023, pp. 108–120. https://doi.org/10.1007/978-3-031-28663-6_9

[14] J. Li *et al.*, "Overview of IEEE BigData 2024 Cup challenges: Suicide ideation detection on social media," in *Proc. 2024 IEEE Int. Conf. Big Data (BigData)*, Los Alamitos, CA, USA: IEEE Computer Society, Dec. 2024, pp. 8532–8540. https://doi.org/10.1109/BigData62323.2024.10825048

[15] S. T. Rabani, A. M. U. D. Khanday, Q. R. Khan, U. A. Hajam, A. S. Imran, and Z. Kastrati, "Detecting suicidality on social media: Machine learning at rescue," *Egypt. Inform. J.*, vol. 24, no. 2, pp. 291–302, 2023. https://doi.org/10.1016/j.eij.2023.04.003

[16] J. Clusmann *et al.*, "The future landscape of large language models in medicine," *Commun. Med.*, vol. 3, p. 141, Oct. 2023. https://doi.org/10.1038/s43856-023-00370-1

[17] D. Alghazzawi, H. Ullah, N. Tabassum, et al., "Explainable AI-based suicidal and non-suicidal ideations detection from social media text with enhanced ensemble technique," *Scientific Reports*, vol. 15, p. 1111, 2025, https://doi.org/10.1038/s41598-024-84275-6

[18] F. Lashgari, M. Pourvahab, A. Sousa, and S. Pais, "SENTINEL-LLM: Suicide ensemble-based text intelligence and natural language evaluation through large language models," in *Proceedings of the IEEE International Conference on Web Research (ICWR '25)*, Tehran, Iran, April 2025, pp. 299-305, https://doi.org/10.1109/ICWR65219.2025.11006176

[19] N. A. John-Henderson, E. J. White, and T. L. Crowder, "Resilience and health in American Indians and Alaska Natives: A scoping review of the literature," *Development and Psychopathology*, vol. 35, no. 5, pp. 2241–2252, 2023, https://doi.org/10.1017/S0954579423000640

[20] H. Zacarias, J. A. L. Marques, V. Felizardo, M. Pourvahab, and N. M. Garcia, "ECG forecasting system based on long short-term memory," *Bioengineering (Basel)*, vol. 11, no. 1, p. 89, Jan. 2024, https://doi.org/10.3390/bioengineering11010089

[21] A. M. Schoene, G. Lacey, A. P. Turner, and N. Dethlefs, "Dilated LSTM with attention for classification of suicide notes," in *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, Hong Kong, 2019, pp. 136–145. https://doi.org/10.18653/v1/D19-6217

[22] S. L. Mirtaheri, S. Greco, and R. Shahbazian, "A self-attention TCN-based model for suicidal ideation detection from social media posts," *Expert Systems with Applications*, vol. 255, 2024, p. 124855. https://doi.org/10.1016/j.eswa.2024.124855

[23] A. Bansal and R. Beniwal, "Sentiment classification on suicide notes using BERT, Bi-LSTM and CNN," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kamand, India, 2024, pp. 1–5. https://doi.org/10.1109/ICCCNT61001.2024.10724321

[24] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "MentalBERT: Publicly available pretrained language models for mental healthcare," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*, Marseille, France, 2022, pp. 7184–7190. https://doi.org/10.5281/zenodo.7464286

[25] N. Nordin, Z. Zainol, M. H. Mohd Noor, and C. L. Fong, "Explainable machine learning models for suicidal behavior prediction," in *Proceedings of the 6th International Conference on Medical and Health Informatics (ICMHI '22)*, Virtual Event, Japan, 2022, pp. 118–123. https://doi.org/10.1145/3545729.3545754

[26] S. Salmi, S. Mérelle, R. Gilissen, R. van der Mei, and S. Bhulai, "The most effective interventions for classification model development to predict chat outcomes based on the conversation content in online suicide prevention chats: Machine learning approach," *JMIR Mental Health*, vol. 11, 2024, e57362. https://doi.org/10.2196/57362

[27] V. Nguyen and C. Pham, "Leveraging large language models for suicide detection on social media with limited

labels," in *Proc. 2024 IEEE Int. Conf. Big Data (BigData)*, Washington, DC, USA, 2024, pp. 8550–8559. https://doi.org/10.1109/BigData62323.2024.10825313

[28] M. Qorich and R. El Ouazzani, "Advanced deep learning and large language models for suicide ideation detection on social media," *Progress in Artificial Intelligence*, vol. 13, pp. 135–147, 2024. https://doi.org/10.1007/s13748-024-00326-z

[29] W. Zeng and Y. Wu, "Attention-based BiLSTM network for social media suicide detection," in *Proceedings of the 2023 4th International Symposium on Artificial Intelligence for Medicine Science (ISAIMS '23)*, Chengdu, China, 2024, pp. 717–721. https://doi.org/10.1145/3644116.3644236

[30] S. Rappai and G. Ramasamy, "Navigating the emotional maze: Understanding adolescent suicidal ideation using CNN-LSTM model," *Intelligent Decision Technologies*, vol. 18, no. 3, 2024, pp. 1797–1811. https://doi.org/10.3233/IDT-240790

[31] N. Badri, F. Kboubi, and A. Habacha, "Combining FastText and GloVe word embedding for offensive and hate speech text detection," *Procedia Computer Science*, vol. 207, pp. 769–778, 2022, https://doi.org/10.1016/j.procs.2022.09.132

[32] T. H. McCoy and R. H. Perlis, "Applying large language models to stratify suicide risk using narrative clinical notes," *Journal of Mood & Anxiety Disorders*, vol. 10, p. 100109, 2025, https://doi.org/10.1016/j.xjmad.2025.100109

[33] S. Lasri, E. H. Nfaoui, and K. Mrizik, "Suicide ideation and risk detection from social media using GPT models," *Journal of Computer Science*, vol. 20, no. 10, pp. 1349–1356, 2024. https://doi.org/10.3844/jcssp.2024.1349.1356

[34] M. Grimland, J. Benatov, H. Yeshayahu, D. Izmaylov, A. Segal, K. Gal, and Y. Levi-Belz, "Predicting suicide risk in real-time crisis hotline chats integrating machine learning with psychological factors: Exploring the black box," *Suicide and Life-Threatening Behavior*, vol. 54, no. 3, 416-424, Feb. 2024. https://doi.org/10.1111/sltb.13056

[35] L. Ren, H. Lin, B. Xu, S. Zhang, L. Yang, and S. Sun, "Depression detection on Reddit with an emotion-based attention network: Algorithm development and validation," *JMIR Medical Informatics*, vol. 9, p. e28754, Jul. 2021, https://doi.org/10.2196/28754

[36] T. Hasan Sakib, M. Ishak, F. F. Jhumu, and M. A. Ali, "Analysis of suicidal tweets from Twitter using ensemble machine learning methods," in *Proceedings of the ACM International Conference on Machine Learning (ACMI)*, Rajshahi, Bangladesh, 2021, pp. 1-7, https://doi.org/10.1109/ACMI53878.2021.9528252

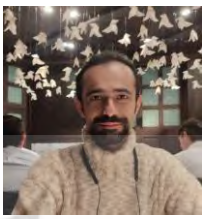
[37] R. C. Cabral, S. C. Han, J. Poon, and G. Nenadic, "MM-EMOG: Multi-label emotion graph representation for mental health classification on social media," *Robotics*, vol. 13, no. 3, p. 53, 2024, https://doi.org/10.3390/robotics13030053

[38] M. Pourvahab *et al.*, "A cluster-based opposition differential evolution algorithm boosted by a local search for ECG signal classification," *Journal of Computational Science*, vol. 86, p. 102541, 2025, https://doi.org/10.1016/j.jocs.2025.102541

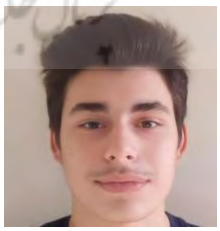
[39] A. Basyouni, H. Abdulkader, W. S. Elkilani, A. Alharbi, Y. Xiao, and A. H. Ali, "A suicidal ideation detection framework on social media using machine learning and genetic algorithms," *IEEE Access*, vol. 12, pp. 124816–124833, 2024, https://doi.org/10.1109/ACCESS.2024.3454796

[40] S. Zad, J. Jimenez, and M. Finlayson, "Hell hath no fury? Correcting bias in the NRC emotion lexicon," *in Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, Aug. 2021, pp. 102–113. https://doi.org/10.18653/v1/2021.woah-1.11

Farzaneh Lashgari is a Ph.D. candidate in Computer Engineering at the University of Beira Interior (UBI), Portugal, and a researcher at HULTIG Lab. Her work focuses on applying large language models (LLMs) to mental health analysis, particularly the early detection of psychological distress and suicidal ideation from social media. She is also a member of NOVA LINCS-UBI, IEEE, and IEEE Women in Engineering (WIE).



Mehran Pourvahab, IEEE Senior Member, is a researcher and educator in computer engineering with broad experience in medical informatics, data science, and network security. He holds a Ph.D. in computer engineering from Azad University, Iran, and is currently an Invited Auxiliary Researcher at the University of Beira Interior (UBI), Portugal. He has completed two postdoctoral fellowships at UBI, contributing to AI-driven healthcare platforms, semantic interoperability, and data modeling in EU-funded projects such as Phara-On. His research also explores the application of large language models (LLMs) and machine learning in mental health. He has served as a faculty member in Iran, supervised postgraduate students, and published extensively in peer-reviewed journals and conferences. His work integrates academic rigor with practical experience in cloud forensics, data analytics, and intelligent systems.



António Sousa is an MSc student in Computer Science at the University of Beira Interior (UBI), Portugal. He is currently a member of the HULTIG Lab (Human Language Technology and Bioinformatics Group). His expertise includes large language models, prompt engineering, empathy modeling, and human-centered AI. He is also interested in affective computing, ethical NLP, and techniques such as zero-shot and few-shot learning. His current research focuses on NLP-driven approaches for mental health and emotion-aware systems.

Anilson Monteiro holds a B.Sc. in Electrical Engineering (Telecommunications) and an M.Sc. in Computer Science, specializing in AI and Machine Learning. His experience spans designing and building full-stack, cloud-native solutions and R&D efforts in software and data engineering across both industry and academia.

Sebastião Pais is a researcher and educator in computer science with multidisciplinary expertise in statistical natural language processing, machine learning, and artificial intelligence. He holds a Ph.D. in Computer Science jointly awarded by MINES ParisTech (France) and the University of Beira Interior (UBI), Portugal. Currently affiliated with the Department of Computer Science at UBI, he began his academic career there in 2016 and has conducted research at internationally recognized institutions, including NOVA LINCS (Portugal) and the CNRS GREYC UMR 6072 Laboratory (France). His work spans a wide range of topics, including lexical semantics, statistical learning, social network analysis, information retrieval, sentiment analysis, and data mining. In 2017, he received the national academic qualification of *Maître de Conférences* (MCF) in Computer Science from the French Ministry of Higher Education, Research, and Innovation. His contributions combine theoretical rigor with applied research across European collaborative projects, promoting innovation in natural language technologies and data science.