Design of a conversational agent to support people on suicide risk

Mario Manso Vázquez, José Manuel Ramírez Sánchez, Carmen García-Mateo,

Laura Docío-Fernández, Manuel José Fernández-Iglesias
atlanTTic Research Center, Universidade de Vigo, Spain.
mario.manso@uvigo.es

Beatriz Gómez-Gómez, Beatriz Pinal, Antia Brañas, Alejandro García-Caballero
Galicia Sur Health Research Institute (IISGS), Spain.
alejandro.alberto.garcia.caballero@sergas.es

Abstract

In this paper, we present a core component of the VisIA project: a conversational agent designed to detect suicide risk factors during real-time chat interactions. By adhering to clinical guidelines and the state-of-the-art theories of suicide, the agent aims to provide a scalable and effective approach to identifying individuals at risk. Preliminary results demonstrate the feasibility and potential of conversational agents in enhancing suicide risk detection.

1 Introduction

Suicide is one of the leading causes of death among young adults worldwide, and its prevention remains a critical public health priority (De Quiroga et al., 2019; WHO, 2019). Current suicide risk assessments methods are often short, resulting in false positives and negatives, and highlighting the need for innovative and scalable approaches (Johnston et al., 2022).

The VisIA Project (Ramírez Sánchez et al., 2024) addresses this challenge by leveraging Artificial Intelligence (AI) technologies (Ji et al., 2020) and multi-modal data, grounded in state-of-the-art theories of suicide (Van Orden et al., 2010; Tsai et al., 2021). The project consists of two major steps: first, conducting a clinical trial to gather clinically validated data, and second, developing solutions to improve suicide risk detection and support systems. At its core is VisIA-Bot, a conversational agent designed to detect suicide ideation during chat interactions. By leveraging suicide constructs, the agent identifies key risk factors and provides targeted support for individuals experiencing emotional distress.

This paper focuses on the VisIA-Bot conversational agent, particularly on the suicide ideation detection component based on suicide prevention theory and practice. The following sections detail the clinical trial design, the VisIA-Bot's suicide constructs detection system and study findings.

2 Clinical Trial

The VisIA Project's clinical trial, see all the details in (Ramírez Sánchez et al., 2024), follows a non-interventional, analytical, observational and prospective design aimed at gathering data from adolescents and young adults (aged 11-16) with varying levels of suicide risk. The study includes a total of **339 participants** divided in three distinct groups: a clinical, a clinical control and a general control populations.

The study has been approved by the Clinical Research Ethics Committee of Galicia (dictum 2023/029), adheres to the Declaration of Helsinki, and the standards of the General Data Protection Regulation (Regulation, 2016). Informed consent was obtained from all participants, and the study is registered under NCT06341634.

3 Theory of Suicide

The understanding of suicide has evolved significantly in recent decades, with contemporary theories emphasizing the interplay of psychological, interpersonal, and experiential factors in the development of suicidal ideation and behaviors. Recent multidimensional frameworks provide greater insight into the complex mechanisms underlying suicide risk. Among these, **Klonsky's Three-Step Theory (3ST)** (Tsai et al., 2021) and **Joiner's Interpersonal Theory of Suicide (ITS)** (Van Orden et al., 2010) have gained prominence for their ability to explain both the emergence of suicidal ideation and the progression to suicide attempts. The ITS posits that suicidal behavior arises from the convergence of two interpersonal constructs, **perceived burdensomeness** (belief of being a liability to others) and **thwarted belongingness** (sense of social disconnection), along with the **acquired capability** for suicide, which develops through habituation to fear and pain via exposure to traumatic or self-injurious experiences. The 3ST, provides a structured framework for understanding suicidal ideation and behavior through three progressive stages: the emergence of suicidal thoughts due to intense **psychological pain** coupled with **hopelessness**; the amplification of suicidal ideation when individuals feel **burdensome** and **disconnected** from others; and the transition to suicide attempts, facilitated by **acquired capability**.

Based on these theories, a practical tool called Suicide Log was presented in (Bryan et al., 2017), aiming at co-constructing and understanding the user's emotional pain through several phases. This tool is commonly used in clinical practice.

4 Related Work

Suicidal ideation detection systems primarily rely on a combination of machine learning, deep learning, and natural language processing (NLP) techniques(Haque et al., 2022; Elsayed et al., 2024). Traditional machine learning models, such as Support Vector Machines (SVM), Random Forest (RF), Decision Trees (DT), and Naïve Bayes (NB), utilize handcrafted feature extraction methods to classify text. More advanced deep learning approaches, such as Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN) and Gated Recurrent Unit (GRU) models, leverage word embeddings to capture semantic relationships within text. These methods work on existing datasets (KOMATI, Accessed: 2023-12-24) for training and testing the models.

The VisIA project follows a different approach, leveraging LLMs to adhere to established clinical practices while integrating insights from modern suicide theories and psychiatry experts.

5 VisIA-Bot Concept

The project proposed the development of a tool based on conversational agent technology designed to follow a structured methodology aimed at supporting clinical practices as well as triage and risk assessment in non-clinical settings, such as school counseling and hospital emergency rooms. The tool aims to identify suicidal constructs and potential suicidal ideation, bridging the gap between early detection and professional intervention. To achieve this, the suicide log was chosen as the primary reference, alongside the theoretical framework guiding its implementation. The interaction begins with open-ended questions focused on the participant's emotions. If self-harm or moderate to high suicide risk is detected, the suicide log procedure is started (see 1).

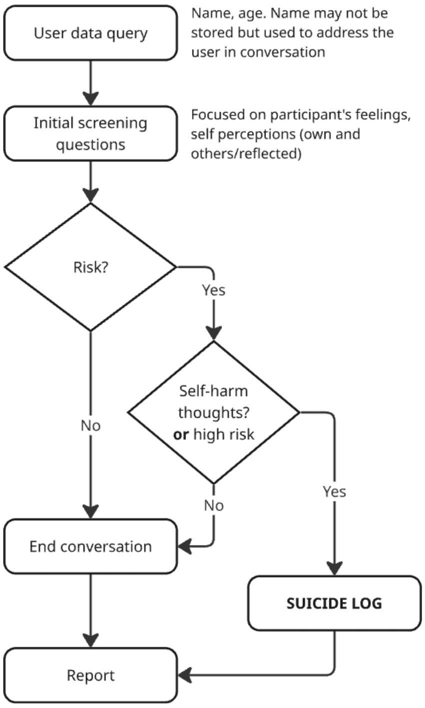

User data query
Name, age. Name may not be stored but used to address the user in conversation
Initial screening questions
Focused on participant's feelings, self perceptions (own and others/reflected)
Risk?
Yes
No
Self-harm thoughts? **or** high risk
No
Yes
End conversation
SUICIDE LOG
Report

Figure 1: Screening basic schema

VisIA-Bot consists of two functionally distinct components that work in parallel: a **conversational agent** designed to follow the principles and steps of the **suicide log**, which guides the conversation, and a **suicide construct detector**, the focus of this paper, designed to identify relevant constructs in text fragments in order to assess suicide risk based on the theoretical framework previously outlined. The interaction between the two components is constant, as the decisions from the conversational agent are based on both the responses of the user and the results of the suicide construct detector.

6 Development Framework

The tool is being developed using LangChain[1] with LangGraph [2] to orchestrate the workflow and Ollama to run the Large Language Models (LLMs). This framework was selected over traditional conversational agent tools like Rasa (Bocklisch et al., 2017) due to its capabilities to generate human-like organic responses, to make decisions based on detailed instructions and to solve complex tasks while simplifying the development process. One of the key limitations of frameworks like Rasa lies

[1] GitHub repo: https://github.com/langchain-ai/langchain
[2] GitHub repo: https://github.com/langchain-ai/langgraph

in their reliance on predefined intents and entities, which can restrict their ability to handle complex or ambiguous inputs. Although traditional frameworks are effective in structured dialogue systems, its rule-based and classification-driven approach struggles with nuanced language, making it less suitable for detecting abstract concepts such as loneliness, distress, or emotional well-being (He and Garner, 2023). Additionally, LLM-based agent orchestration capabilities provide a clear advantage in decision-making and task-solving (Shen et al., 2023).

Regarding LLMs, *Llama3:8B* and *Mistral:7B* were considered for development among other optimized models, being *Llama3* the best performer in early testing and final option. Both models were queried to assess their knowledge of the theoretical framework, confirming their familiarity with relevant concepts and their interrelations.

7 Suicide detection using LLMs

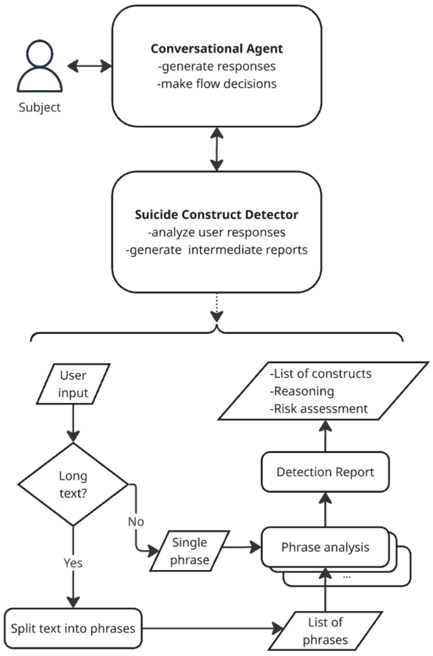


Figure 2: VisIA-Bot Overview

7.1 Definition of Suicide Constructs to Target

The focus was set on the constructs extracted directly from the theoretical framework, following an integrative approach, since both theories have distinct perspectives on critical constructs. According to Klonsky, *"disrupted connectedness is similar to low belongingness and burdensomeness as described in Joiner's Interpersonal Theory"*. For this work, since low belongingness is conceptually very close to disrupted connectedness and would be potentially very difficult to discern, both were integrated into one construct. During development, the construct of acquired capability yielded consistently low detection rates, leading to its replacement with two constructs that could be identified more clearly in adolescent's statements: passive suicidal ideation and active suicidal ideation. In these new constructs, the acquired capability is implicitly included within active suicidal ideation, which represents the higher-risk construct. According to Klonsky's theory, suicidal ideation arises from psychological pain and hopelessness, which can be directly expressed in concise chat phrases, such as "I wish I could fall asleep and never wake up". In conclusion, the final **suicide constructs** to be detected are: psychological pain, hopelessness, burdensomeness, disrupted connectedness (low belongingness), passive suicidal ideation and active suicidal ideation.

7.2 Suicide Construct Detection Strategies

The system is designed to detect suicide-related constructs in short to medium-length phrases, prioritizing real-time responsiveness. After evaluating multiple detection strategies, the most effective and practical approach was selected due to its simplicity, low computational overhead, and alignment with real-time system requirements: analyzing individual, context-free phrases. Additionally, the absence of robust real-word datasets for contextual approaches reinforces the choice of single-phrase analysis.

To address the limitation of short-phrase analysis while maintaining real-time efficiency, the following best practices are proposed:

- Segmented Text Analysis: Dividing longer texts into smaller fragments for individual analysis, computing aggregated results, can improve computational efficiency and is aligned with clinical practices.

- State Variables for Suicide Constructs: Introducing state variables allows for dynamic tracking of suicidal ideation throughout a conversation, enabling its inference based on construct concurrence. This minimizes dependence on explicit indicators of suicidal ideation (e.g., the concurrence of psychological pain and hopelessness), allowing for a more comprehensive and nuanced assessment of the user's psychological state.

- Multi-Tiered Analysis: Implementing a lightweight prefiltering model to identify potentially relevant phrases before LLM processing minimizes computational overhead while enhancing accuracy.

- Structured Output: Possible hallucinations were controlled by forcing a structured output and ensuring the context window of the model is never exceeded.

7.3 Prompting strategies

Combinations of three main prompting strategies (Wang et al., 2023) were explored for the detection of suicide constructs in text: prompt engineering, few shot and Retrieval Augmented Generation (RAG). In all cases, the theoretical framework was contextualized in the prompt. These strategies included:

- instruction-based prompting (IP), leveraging the LLM's internal knowledge,

- few-shot prompting (IP+FS), providing a curated list of example phrases associated with each construct,

- Retrieval-Augmented Generation (RAG), adding theoretical context (RAG).

The IP prompting strategy was developed using prompt engineering and provides clear instructions to the LLM, but no examples of phrases. The prompt instructions were developed iteratively using a test set generated with GPT-o4[3] and selected fragments of text from the first stages of the clinical trial, analyzing its reasoning for each detection. The IP+FS strategy, based on few shot prompting, provides both clear instructions and between 10 to 20 example phrases for each construct. These examples were crafted by psychiatry experts focusing on variety, trying to maximize case coverage while minimizing overlap and maintaining an equitable number of entries per construct to prevent bias. The RAG approach is based on the IP prompt adding a context retrieved from a knowledge base which contains the theoretical framework and detailed descriptions and examples for each construct.

The objective is to identify suicide-related constructs within a set of 80 test phrases. This set consists of 30 neutral phrases and 50 phrases associated with suicide constructs, with 8 per construct, except for psychological pain and disconnection, which have 9 each. Since, to the best of our knowledge, no dataset of phrases associated with these constructs is available, the test phrases were generated based on psychiatrists' instructions and subsequently reviewed by them, following the same criteria of the prompt examples regarding coverage and overlap, while ensuring no overlap or redundancy between sets. The order of the test set is randomized on each run to prevent model bias. Prompting strategies were also tested on real clinical trial texts.

Detection outputs are structured as a JSON object with **emotion**, **confidence**, and **reasoning** fields. The LLM selects a single construct when multiple are detected and provides reasoning to explain its decision. This reasoning component is essential for understanding the model's decision-making process and identifying factors contributing to detection success or failure. Here is a result example for the sentence "*At times, I am overwhelmed by the idea of disappearing, but I don't know how or when it might happen.*" classified by the clinical team as passive suicidal ideation:

```
{'emotion': 'passive suicidal ideation',
'reasoning': 'The sentence indicates
    that the user feels the idea of
    disappearing, which suggests
    possible passive suicidal ideation.
    The lack of specificity about how or
     when it might happen does not rule
    out this possibility.'}
```

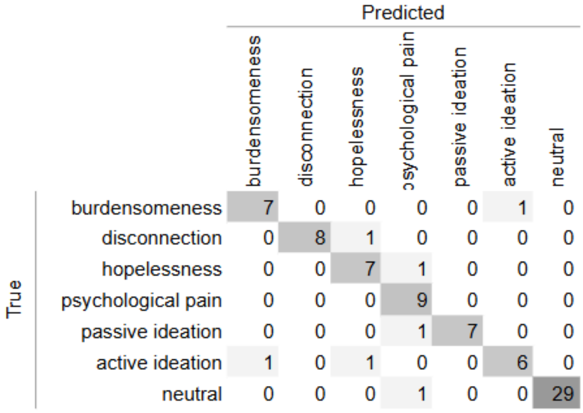
The results were analyzed across several dimensions, including accuracy for overall performance and by category (suicide constructs vs. neutral phrases), confusion matrix, Precision, Recall and F1-score. The results for one run of the IP-FS strategy,, which achieved the best performance overall, are shown in Table 1 and Fig. 3. In this run, the overall accuracy was , 90% and the category accuracy was 98%. The LLM used was *Llama3*.

Labels	Precision	Recall	F1-Score
burdensomenss	88%	88%	88%
disconnection	100%	89%	94%
hopelessness	78%	88%	82%
psychological pain	69%	100%	82%
passive ideation	100%	75%	86%
active ideation	86%	75%	80%
neutral	100%	97%	98%

Table 1: Evaluation metrics for the IP+FS strategy

The comparative results of the three prompting

[3] Hello GPT-4o | OpenAI



		Predicted						
		burdensomeness	disconnection	hopelessness	psychological pain	passive ideation	active ideation	neutral
True	burdensomeness	7	0	0	0	0	1	0
	disconnection	0	8	1	0	0	0	0
	hopelessness	0	0	7	1	0	0	0
	psychological pain	0	0	0	9	0	0	0
	passive ideation	0	0	0	1	7	0	0
	active ideation	1	0	1	0	0	6	0
	neutral	0	0	0	1	0	0	29

Figure 3: Confusion matrix for the IP-FS strategy.

strategies for Precision, Recall and F1-Score macro average and weighted average are shown in Fig. 4 and Fig. 5 respectively. RAG yielded the worst performance, getting a Precision macro average of 68%, followed by IP, which reached 77%. The best results were achieved using IP+FS, getting 0.89% in this particular run. The average values for 10 runs are similar to this result: Precision macro average: 0.89; Recall macro average: 0.87; Precision weighted average: 0.91; and Recall weighted average: 0.90.

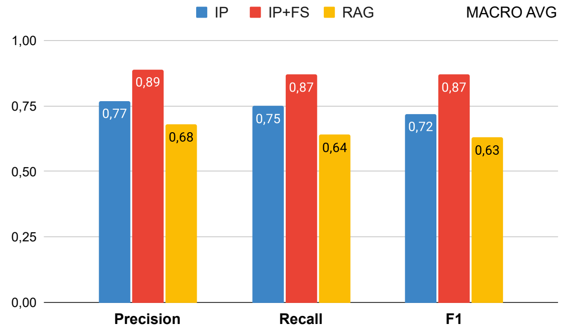


Figure 4: Precision, Recall and F1 Macro Average

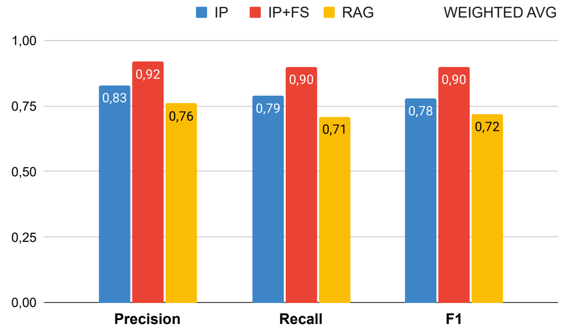


Figure 5: Precision, Recall and F1 Weighted Average

8 Discussion and Future Work

This paper establishes short-phrases analysis as the primary strategy for detecting suicide constructs, balancing real-time efficiency with expert-in-the-loop refinement, achieving ≅90% accuracy, a weighted average for Precission and Recall of 0.92 and 0.90 respectively, despite challenges in overlapping constructs. Hopelessness and psychological pain were the constructs with lower precision values, with 0.69 and 0.78 respectively, but with higher recall values, with 0.88 and 1. It was observed that some of the test phrases regarding these constructs exhibit significant variability in classification between them This may indicate an overlap in the semantic or emotional representation of these constructs, or it may reflect differences in how the model interprets subtle linguistic cues. Further research with real data from the clinical trial is planned to improve these results.

To further validate the results of the test set, the same evaluation previously performed by the model will be performed by medical professionals, specifically psychologists and psychiatrists. This comparison seeks to evaluate the concordance between the model's predictions and expert assessments while enhancing the test's reliability and construct validity.

The detection of suicide constructs in long texts is under development, focusing on text segmentation and state variables to analyze construct combinations and repetitions. Multi-tiered analysis is being explored for non-real-time scenarios to optimize construct detection through effective sentence division.

Additionally, recent findings emphasize the importance of incorporating positive constructs, such as protective factors like connectedness and emotional granularity, into suicide risk assessment. These factors, even when contradicting other risk indicators, directly influence risk evaluations, as demonstrated in clinical trial transcripts where both risk and protective factors co-occurred.

Acknowledgments

VisIA project is funded by the 2021 call for strategic projects aimed at the ecological transition and the digital transition, of the Spanish plan for scientific, technical and innovation research 2021-2023, within the framework of the recovery, transformation and resilience plan (References: TED2021-130747B-C21 and TED2021-130747B-C22).

References

Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. 2017. Rasa: Open Source Language Understanding and Dialogue Management. *arXiv e-prints*, arXiv:1712.05181.

Craig J Bryan, Jim Mintz, Tracy A Clemans, Bruce Leeson, T Scott Burch, Sean R Williams, Emily Maney, and M David Rudd. 2017. Effect of crisis response planning vs. contracts for safety on suicide risk in us army soldiers: A randomized clinical trial. *Journal of affective disorders*, 212:64–72.

S De Quiroga, M Riesgo, E Martín del Campo, S Pulido, and S Rodrigo. 2019. Impacto socioeconómico de la depresión y el suicidio en españa. *Rev Esp Econ Salud*, 14(5):923–47.

Nelly Elsayed, Zag ElSayed, and Murat Ozer. 2024. Cautionsuicide: A deep learning based approach for detecting suicidal ideation in real time chatbot conversation. In *2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI)*, pages 1–5. IEEE.

Rezaul Haque, Rezaul Haque, Rezaul Haque, Naimul Islam, Naimul Islam, Naimul Islam, Maidul Islam, Maidul Islam, Maidul Islam, Md Manjurul Ahsan, and Manjurul Ahsan. 2022. A comparative analysis on suicidal ideation detection using nlp, machine, and deep learning. *Technologies (Basel)*.

Mutian He and Philip N. Garner. 2023. Can chatgpt detect intent? evaluating large language models for spoken language understanding. *Preprint*, arXiv:2305.13512.

Shaoxiong Ji, Shirui Pan, Xue Li, Erik Cambria, Guodong Long, and Zi Huang. 2020. Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Transactions on Computational Social Systems*, 8(1):214–226.

Jenessa N Johnston, Darcy Campbell, Hector J Caruncho, Ioline D Henter, Elizabeth D Ballard, and Carlos A Zarate Jr. 2022. Suicide biomarkers to predict risk, classify diagnostic subtypes, and identify novel therapeutic targets: 5 years of promising research. *International journal of neuropsychopharmacology*, 25(3):197–214.

N. KOMATI. Accessed: 2023-12-24. Suicide and depression detection.

José Manuel Ramírez Sánchez, Mario Manso, Carmen García-Mateo, Beatriz Gómez-Gómez, Beatriz Pinal, Antía Brañas, Alejandro García Caballero, Laura Docío-Fernandez, and MJ Fernández-Iglesias. 2024. Visia project: design of an automated ai-based emotional distress and suicide risk detection system. In *Proc. IberSPEECH 2024*, pages 275–277.

Protection Regulation. 2016. Regulation (eu) 2016/679 of the european parliament and of the council. *Regulation (eu)*, 679:2016.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Preprint*, arXiv:2303.17580.

Michelle Tsai, Harris Lari, Samantha Saffy, and E David Klonsky. 2021. Examining the three-step theory (3st) of suicide in a prospective study of adult psychiatric inpatients. *Behavior therapy*, 52(3):673–685.

Kimberly A Van Orden, Tracy K Witte, Kelly C Cukrowicz, Scott R Braithwaite, Edward A Selby, and Thomas E Joiner Jr. 2010. The interpersonal theory of suicide. *Psychological review*, 117(2):575.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

World Health Organization WHO. 2019. Suicide in the world: global health estimates. Technical report, World Health Organization.