ISBN: 978-628-95207-8-1. ISSN: 2414-6390. Digital Object Identifier: 10.18687/LACCEI2024.1.1.1547

Maturity Model Based on a Suitability Approach for the Evaluation of Chatbots Used in Depression Detection

Fernando David Mori Muñoz, Bachelor of Information Systems Engineering[1], Rodrigo Alonso Berrocal, Bachelor of Information Systems Engineering[2], Edgar David Diaz Amaya, Magister[3]
[1,2,3]Peruvian University of Applied Sciences, Perú, u20191a700@upc.edu.pe, u201816249@upc.edu.pe, pcsiedia@upc.edu.pe

Abstract- In the field of psychology, the use of artificial intelligence-based depression detection chatbots is being employed in order to reduce the percentage of people with depression in the world. However, 3 out of 8 sessions conducted to these software products are not completed due to lack of confidence or self-esteem, trustworthiness and safety of the user. This is due to the disengagement of the chatbot in the conversation it holds with users and the color connotation employed. To avoid producing chatbots with this quality, this research presents a maturity model to evaluate these conversational agents, combining a questionnaire to measure the usability of mobile health applications, a performance metric to measure the chatbot's ability to detect depression, and a proposed category that evaluates whether appropriate depression detection tools were used when training the classification model to detect depression; the results obtained indicate that this model could achieve an average additional accuracy of 6% when evaluating a chatbot.

Keywords--maturity model, artificial intelligence, software product, chatbot and "depression.

I. INTRODUCTION

The World Health Organization mentions that 3.8% of the world's population suffers from depression, that is, approximately 280 million people [21]. In view of this, different technological companies are venturing into the medical field using emerging technologies in order to contribute and minimize the percentage of people suffering from this mental disorder [11]. Chatbots have been developed for the detection of depression based on artificial intelligence. However, 3 out of 8 sessions performed to these software products are not completed [9]. This raises a big question about the factors that cause users not to complete their sessions. According to [9], the declarative psychological responses by these bots bring significant negative impacts to the user such as lack of confidence or self-esteem as a consequence of the chatbot's disengagement from the conversation. In addition, the color connotation employed does not reflect the trustworthiness or security that the user requires [14], concluding that there is no process-oriented evaluation method for this type of software [16].

According to [16], evaluating the different characteristics of a chatbot can be a difficult challenge, because the artificial intelligence module of the software product is usually improved. However, these usually have a focus only on the artificial intelligence algorithms and do not help to improve the quality attributes.

Our goal is to find a tool that can evaluate the effectiveness and usability of chatbots, since they provide the depressed user with a company with whom he can interact and communicate his problems, anxieties and thoughts so that it can provide psychological support similar to that of a psychologist to counteract the severity of these mental disorders.

Therefore, a maturity model was developed consisting of the categories of suitability, usability and performance, which attempt to measure the chatbot in terms of usability and effectiveness. To validate its contribution to the solution of the problem, experiments were conducted with the help of an artificial intelligence-based depression detection chatbot. These experiments allowed us to achieve an increase in accuracy of 6% compared to other ways of evaluating this type of chatbots as detailed below.

II. RELATED WORKS

The field of chatbot evaluation has become a crucial area of research, as determining their quality and accuracy in response and experience that can have a significant impact on the user experience and the success of their implementation. This section mentions related work that uses techniques similar to those used by us. These researches mention proposals or alternative approaches to improve aspects related to the evaluation, usability or satisfaction of different technologies in the health field.

From the point of view of other research on evaluation metrics that we can link to software products related to the healthcare sector, a maturity model for evaluating and improving software testing practices called TMMi has been proposed in [10]. This model provides an overview of global software testing trends and certifications and is based on a large number of metrics from the QA sector. At [22], a similar approach was proposed, as the author adapts an agile testing maturity model to integrate it to software products to improve their quality and efficiency. Also, a group conformed by five usability heuristics was proposed for the improvement of these chatbots, but experts made observations about the effectiveness and clarity of the solution [16]. The author in [18] raised a similar approach, because he identified which are the most important features to evaluate the quality in use based on ISO/IEC25010 for clinical software. In addition to this, the author in [03] proposed a framework called QinUF to measure software quality in use based on semantic similarity and sentiment analysis. That is, it determines the quality of a software in use based on its sentences and score.

In [1], the authors propose that the sociocultural and behavioral aspects of mHealth users should be considered when designing and implementing mobile health applications. They also suggest improving the functions and features of such apps

to more effectively meet users' needs, thereby increasing their satisfaction and willingness to use them in the future. In contrast, we propose functionality enhancements to improve effectiveness, usability, and satisfaction.

In [4], the authors propose an improved methodology to evaluate the quality in use of clinical chatbots. This methodology is based on the ISO/IEC 25010 standard and uses the AHP method to compare different clinical chatbots based on their quality in use. The methodology also includes an additional contribution by defining a quantitative method for calculating AHP pairwise comparisons. In contrast, in our model we employ the ISO/IEC 25040 standard, which focuses on measuring and evaluating the quality in use of the software and meeting the needs of users to achieve specific objectives in a given context.

In [14], the authors propose a set of five usability heuristics for evaluating chatbots, which are based on experience in the development of this type of applications and on a thorough review of the state of the art. These heuristics were tested in a case study with the help of five experts, who evaluated an education-oriented chatbot. The results revealed that, although the proposed heuristics need to be refined, they are an excellent first step to broaden the horizon of usability evaluations in chatbots. On our side, we propose quantitative heuristics to measure the effectiveness, usability in chatbots aimed at depression detection and evaluation.

In [24], the authors propose the development and validation of a new mobile health app usability questionnaire (MAUQ) to assess the usability of mobile health apps. The study was based on existing questionnaires used in previous mobile application usability studies. Our research work proposes a maturity model to measure not only usability, but also tool performance and suitability, based on questionnaires.

In [15], the authors propose a machine learning-based classification model for the detection of coronary heart disease, which uses clinical data and expert opinion as input to predict the presence or absence of coronary heart disease in patients. In addition, they propose the use of metrics such as accuracy, sensitivity, specificity, Jaccard score, F1-score and confusion matrix to assess the accuracy of the model. In contrast, we propose metrics such as usability and performance as evaluation in order to improve the usability and effectiveness of the chatbot.

In [15], the authors propose a machine learning-based classification model for the detection of coronary heart disease, which uses clinical data and expert opinion as input to predict the presence or absence of coronary heart disease in patients. In addition, they propose the use of metrics such as accuracy, sensitivity, specificity, Jaccard score, F1-score and confusion matrix to assess the accuracy of the model. In contrast, we propose metrics such as usability and performance as evaluation in order to improve the usability and effectiveness of the chatbot.

III. MATURITY MODEL TO EVALUATE CHATBOTS USED FOR DEPRESSION DETECTION

A. Preliminary concepts

Definition 1 (Maturity Model [7]*):* A structured framework used to assess an organization's level of competence and performance in a specific area by identifying critical processes and defining best practices at each stage of that area's life cycle, enabling the organization to evaluate its current performance, identify areas for improvement, and establish a clear path to achieve a higher level of excellence.

Definition 2 (ISO/IEC 25040 [13]*):* It is a worldwide standard that establishes guidelines for assessing software and system quality, as an integral part of the series dedicated to requirements and quality assessment of systems and software. This standard defines the process for software product evaluation. This process has the activity of "Establish the evaluation requirements", "Specify the evaluation", "Design the evaluation", "Execute the evaluation" and conclude the evaluation".

Definition 3 (mHealth App Usability Questionnaire [24]*):* It is an instrument designed and validated to evaluate the usability of health applications, by means of a questionnaire that evaluates aspects such as ease of use, system layout and usefulness of the application using a 7-point Likert scale. This scale has the option of "Strongly disagree", "Disagree", "Moderately disagree", "Neither agree nor disagree", "Moderately agree", "Agree" and "Strongly agree".

Definition 3 (mHealth App Usability Questionnaire [24]*):* It is an instrument designed and validated to evaluate the usability of health applications, by means of a questionnaire that evaluates aspects such as ease of use, system layout and usefulness of the application using a 7-point Likert scale. This scale has the option of "Strongly disagree", "Disagree", "Moderately disagree", "Neither agree nor disagree", "Moderately agree", "Agree" and "Strongly agree".

Definition 4 (Area under the receiver operating characteristic curve [8]*):* AUC-ROC reflects the ability of a classification model to differentiate positive and negative classes using the plot of the true positive rate on the ordinate axis and false positive rate on the abscissa axis. If the AUC-ROC tends to 100%, then the model's capability is considered good. If it tends to 0%, then its capability is considered poor.

Definition 5 (Test Maturity Model Integration [10]*):* It is a structured framework based on best practices, which is used to assess the maturity level of an organization's software testing processes, through the definition of a set of key areas, practices and improvement objectives, in order to improve the quality and effectiveness of software testing and, ultimately, the quality of the software product. The TMMI model aims to provide clear guidance for organizations that want to improve their software testing processes and reach a higher level of maturity in this critical area. In addition, this model is so large that it has a guide for evaluating software products.

Definition 6 (Artificial intelligence-based ChatBot [23]*):* A software product designed to interact with users in a conversational manner, simulating a human conversation. It uses artificial intelligence techniques, such as natural language processing and machine learning, to understand and respond to user questions and requests. These ChatBots can be used in a variety of applications, such as customer service, technical support, web browsing assistance and more, providing fast and personalized responses through chatbot interfaces.

Concept 1 (TMMI and ISO/IEC 25040): The union of Test Maturity Model Integration (TMMI) and ISO/IEC 25040, provides a comprehensive approach towards the assessment of chatbots used in depression detection. On the ISO/IEC 25040 side, it provides an evaluation process that provides guidance on how to select metrics and apply decision criteria. On the TMMI side, this model provides guidance on how to set the quantitative evaluation objectives, how to measure quantitatively, obtain the results to compare with the established objectives and determine the maturity level.

Definition 7 (Depression screening tools [5]*):* Depression screening tools are instruments used to assess symptoms of depression in an individual. These questionnaires or scales can be useful in identifying individuals who may be experiencing symptoms of depression and need additional medical or psychological care. The tools to be used depend on the target audience and the area of depressive symptoms assessed by the chatbot. These tools will be described below indicating the name, abbreviation, utility and source:

Table 1. Depression screening tools

Name	***Abbreviation***	***Utility***	***Source***
Beck Depression Scale	BDI	Assesses symptoms of depression based on the emotions and thoughts associated with it. (Durankuş & Aksu, 2022)..	https://pubmed.ncbi.nlm.nih.gov/32419558/
Hamilton Depression Inventory	HDRS	It assesses symptoms of depression based on a variety of domains, including emotional symptoms as well as physical and behavioral symptoms (Braunschneider et al., 2021)..	https://pubmed.ncbi.nlm.nih.gov/33249538/
Geriatric Depression Scale	GDS	Assesses symptoms of depression based on a combination of emotions and physical symptoms related to depression in older adults. (Thompson & Jones, 2020).	https://pubmed.ncbi.nlm.nih.gov/33015310/
Edinburgh Postnatal Depression Scale	EPDS	It assesses symptoms of depression based on a combination of emotions and thoughts related to the postpartum period (Lutkiewicz et al., 2020)..	https://pubmed.ncbi.nlm.nih.gov/32731490/
Patient Health Questionnaire	PHQ-9	It assesses symptoms of depression based on a combination of emotions, thoughts, and physical symptoms related to depression with a 2-week scope. (Burchert et al., 2021)..	https://pubmed.ncbi.nlm.nih.gov/33406120/
Childhood Depression Inventory	CDI-2	Assesses depressive symptoms based on a combination of emotions, cognitions, and depression-related behaviors in children and adolescents (Gijzen et al., 2021)..	https://pubmed.ncbi.nlm.nih.gov/32956963/

B. Method

In this article, we present an approach based on TMMI and ISO/IEC 25040 to evaluate chatbots focused on depression detection. Our maturity model seeks to provide a clear and organized structure to systematically assess an organization's capability in depression detection using chatbots, identifying both strengths and weaknesses.

1) Model

The main contribution of this paper is the design of a model that proposes three categories to evaluate chatbots. The first category is called "Suitability of depression detection tools". This category evaluates whether appropriate depression detection tools were used when training the model to detect depression. That is, whether the datasets used to train the artificial intelligence module implemented in the depression detection chatbot were appropriate taking into account the scope of the target audience of this conversational agent. Table 4 below shows the structure of the calculation of this category.

Table 2. Description of the questions related to the category "tool suitability".

Ask	***Depression screening tool***
Does the chatbot detect depression based on emotions and the thoughts associated with it?	When training the classification model, did you use data sets related to the Beck Depression Scale?
Does the chatbot detect depression based on emotional symptoms as well as physical and behavioral symptoms?	When training the classification model, did you use data sets related to the Hamilton Depression Inventory?
Does the chatbot detect depression based on emotions and physical symptoms related to depression in older adults?	When training the classification model, did you use data sets related to the Geriatric Depression Scale?
Does the chatbot detect depression based on combination of emotions and thoughts related to the postpartum period?	When training the classification model, did you use data sets related to the Edinburgh Postnatal Depression Scale?
Does the chatbot detect depression based on thoughts and physical symptoms related to depression with a 2-week range?	When training the classification model, did you use data sets related to the Patient Health Questionnaire?
Does the chatbot detect depression based on emotions, cognitions and depression-related behaviors in children and adolescents?	When training the classification model, did you use data sets related to the Childhood Depression Inventory?

Table 2 shows the questions, in the "Question" column and the "Depression detection tool" column, that the evaluator should answer with "Yes" or "No". For example, if the evaluator answers with "Yes" to the question "Does the chatbot detect depression based on emotions and the thoughts associated with it?", then the question in the same row, but in the "Depression detection tool" column, should be used, otherwise the next question in the "Question" column will be asked. In the case of a "Yes" answer to the question above, the question "When training the classification model, did you use data sets related to the Beck Depression Scale?" should be used. This category is calculated with the following formula:

$$\frac{Total\ number\ of\ tools\ to\ be\ implemented}{Total\ tools\ implemented} * 100\% = \%Adequacy\ of\ tools \quad (1)$$

The second category is called "Usability", and the mHealth App Usability Questionnaire is used to measure it. This

questionnaire measures intention to use, ease of use, system layout, usability and satisfaction.

Table 3. mHealth App Usability Questionnaire (MAUQ)[21]

Dimension	mHealth App Usability Questionnaire	
	ID	***Indication***
Ease of use	S1	The application was easy to use.
	S2	It was easy for me to learn how to use the application.
	S3	Navigation was consistent when moving between screens.
	S4	The application interface allowed me to use all the functions.
	S5	Whenever I made a mistake while using the application, I could recover easily and quickly.
Intention to use and satisfaction	S6	I like the interface of the application.
	S7	The information in the application was well organized, so I could easily find the information I needed.
	S8	The application recognized and provided adequate information to inform me of the progress of my action.
	S9	I feel comfortable using this application in social settings.
	S10	The amount of time involved in using this application has been adequate for me.
	S11	I would use this application again.
	S12	Overall, I am satisfied with this application
Utility and system layout	S13	The application would be useful for my health and well-being.
	S14	The application improved my access to health care services.
	S15	The application helped me manage my health effectively.
	S16	This application has all the functions and capabilities I expected it to have.
	S17	I could use the application even when the Internet connection was poor or unavailable.
	S18	This mobile health application provided an acceptable way to receive health care services, such as accessing educational materials, tracking my own activities, and performing self-assessments.

The third category is called "Performance". This category evaluates the performance of the predictive model implemented in the chatbot using the confusion matrix to obtain sensitivity and specificity. In this way, the AUC-ROC can be calculated and the chatbot's ability to detect depression can be determined. Once the result of the 3 categories is obtained, a simple average will be performed to calculate the quality index. The quality index will be calculated with the following formula:

$$Quality\ index = \frac{\%Tool\ suitability + Usability + Performance}{3} \quad (2)$$

This model seeks to evaluate, in an accurate way, a depression detection chatbot by means of the combination of these 3 categories, using a simple average of the results obtained by each one of them. The 5 levels of this model will be described below:

Table 4. Description of the maturity model levels by rank

Level	Range	Comment
1	[0%, 25%>	At this level, the chatbot is difficult to use, due to the lack of an intuitive interface and limited interaction options. This agent detects depression wrongly and focuses on providing general answers and has been trained without taking into account its target audience.
2	[25%, 50%>	At this level, the chatbot is considered moderately difficult to use, unhelpful and unable to satisfy the user. This agent detects depression accurately on certain occasions and has been trained by taking into account a small part of its target audience.
3	[50%, 70%>	At this level, the chatbot is considered neither good nor bad in terms of ease of use, usability and user satisfaction. This agent detects depression accurately almost 50% of the time and has been trained taking into account almost half of its target audience.
4	[70%, 90%>	At this level, the chatbot is considered moderately good in terms of ease of use, usability and user satisfaction. This agent detects depression accurately almost 70% of the time and has been trained with a large part of its target audience in mind.
5	[90%, 100%]	At this level, the chatbot is easy to use, well laid out, helpful and satisfies the user. This agent detects depression accurately almost 90% of the time and has been trained with a large part of its target audience in mind.

2) Process

The second contribution of this paper is to make a comprehensive approach to bring together Test Maturity Model Integration (TMMI) and ISO/IEC 25040 to evaluate chatbots used in depression screening. On the ISO/IEC 25040 side, this provides an evaluation process that provides guidance on how to select metrics and apply decision criteria. On the TMMI side, this model provides guidance on how to establish the quantitative assessment objectives, how to measure quantitatively, obtain the results to compare with the established objectives and determine the maturity level. To apply this model, an assessment process was designed based on the model and standard described above. This process is used when evaluating the 3 dimensions formulated by the model proposed in this work.

- The process starts when the user wants to evaluate a depression screening chatbot.
- As step two, you have to select the categories of the model you want to measure.

- As step three, quantitative evaluation objectives must be established using the model to determine whether these objectives were achieved.
- As step four, measurements have to be performed quantitatively in order to classify the chatbot.
- As step five, the evaluation decision criteria must be applied to determine the result obtained by each selected category.
- Step six is to obtain the results for each category selected.
- As step seven, it is necessary to determine if all the categories of the model were selected in order to calculate the quality index of the chatbot, otherwise the results will be compared with the established objectives of the selected categories.
- As step eight, the quality index has to be calculated in order to determine the level of maturity achieved by the chatbot.

IV. EXPERIMENTS

In this section, we will discuss the experiments of our project. Also, we will describe what is needed to replicate these experiments and discuss the results.

A. Experimental Protocol

The objective of the experiment was to identify the covered criteria of our model with respect to MAUQ and AUC-ROC in order to empirically analyze the variation of our results, through monthly iterations. In this regard, we attempted to describe in detail how different variables affect the final result to gain a deeper understanding of the variations among these. Careful analysis of these criteria not only enriches our knowledge of model performance, but also reveals areas that may require improvement and optimization for future development.

The experimental procedure started with the identification of a depression detection chatbot called "Sequitur". This is a chatbot based on Machine Learning and validated by 18 medical students and 3 qualified psychologists. It is worth noting that these 3 psychologists agreed on the usefulness of the chatbot. There were 3 different versions of this chatbot to evaluate each one of them monthly.

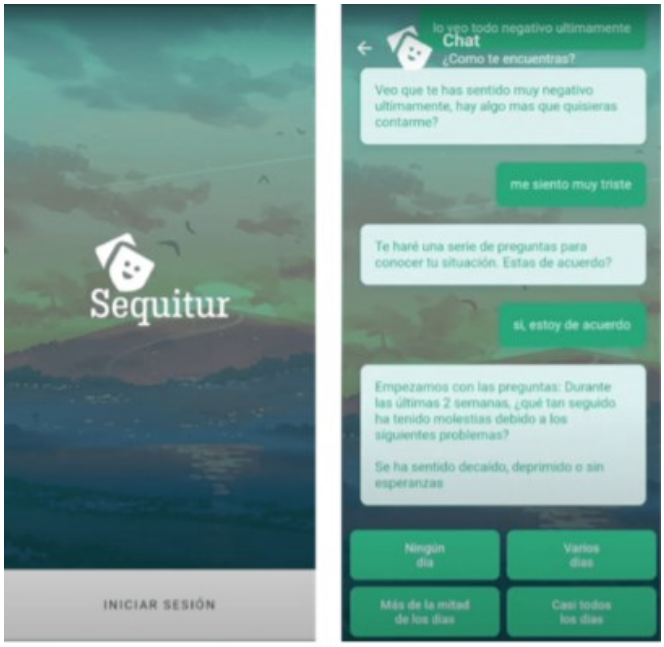

Fig. 1 Sequitur screens

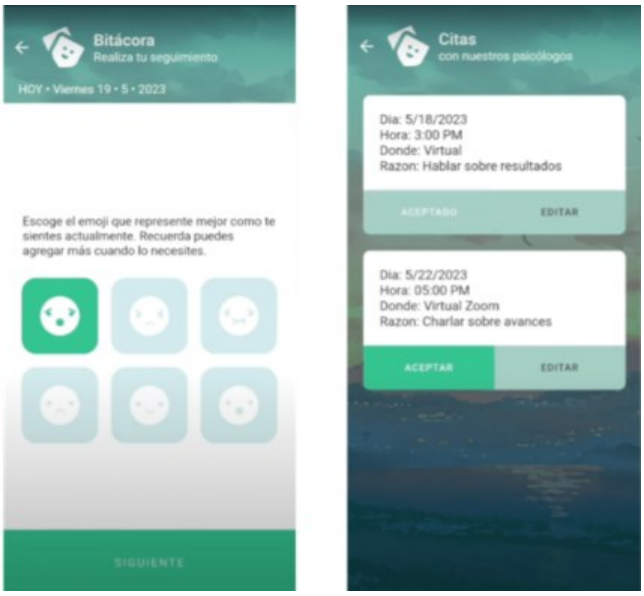

Fig. 2 Recommendation and citation screens

These monthly iterations were carried out from June 2023 to August 2023. For the execution of each iteration, meetings were coordinated with the developers to have knowledge about the changes made between each iteration and clarity at the time of evaluating the chatbot.

B. Results

In the first iteration, the questions of the proposed maturity model were used in their entirety and a percentage of tool suitability equal to 67%, usability equal to 88% and performance equal to 72% was obtained, giving as a final result a quality index equal to 76%. If we look at table 4, we can determine that the maturity level of the chatbot is 4.

In the second iteration, improvements were made to the training of the classification model implemented in the chatbot. The dataset already managed for training the model was improved resulting in 75% performance. Regarding the usability and suitability of detection tools, they continue with the same percentages as iteration 1, since no significant change was evidenced. Therefore, a quality index equal to 77% was obtained. In other words, the maturity level calculated by the maturity model is still 4.

In the third iteration, improvements were made to the training of the classification model and user experience. A dataset related to the Hamilton Depression Inventory was added, which serves to detect depression based on emotional symptoms as well as physical and behavioral symptoms. In terms of user experience, they added new functionalities such as quotes and recommendations that psychologists can provide by reading the results of the severity of depression obtained by patients. It was calculated that the suitability of depression detection tools was equal to 100%, the usability equal to 89% and the performance equal to 82%, obtaining a quality index of 90% and a maturity level of 5:

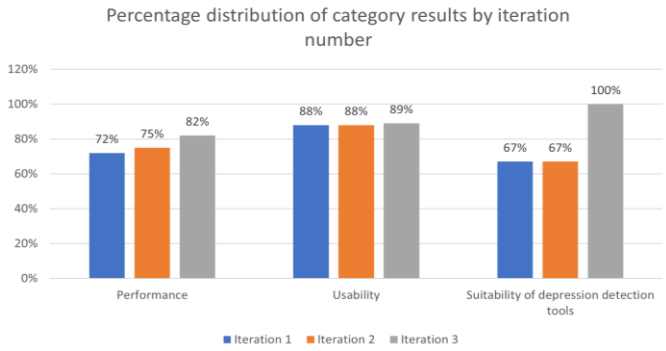

Fig. 3 Bar chart of the results of each category of the maturity model.

The author's approach in [24] allows measuring the usability of a chatbot in terms of intention to use, ease of use, system layout, usefulness and satisfaction. Thanks to these metrics, it was possible to have evaluations sensitive to chatbot design changes. Regarding the chatbot's ability to detect depression, the author's research in. [8] helps us to determine the performance of such a chatbot classification model. This is reflected in Figure 4, since, if we average out the differences of the result of the maturity model, AUC-ROC and MAUQ, we will obtain that in the first iteration there was an average difference of 8%. As for iteration 2, there was an average of 7% and in iteration 3 there was an average of 5%. If we calculate the simple average of the averages of the 3 iterations, we obtain that the maturity model achieves by reaching an average of 6%. These calculations are shown in Table 5. Below are the iterations performed using the 2 approaches separately and unifying these approaches with our Suitability category for our maturity model.

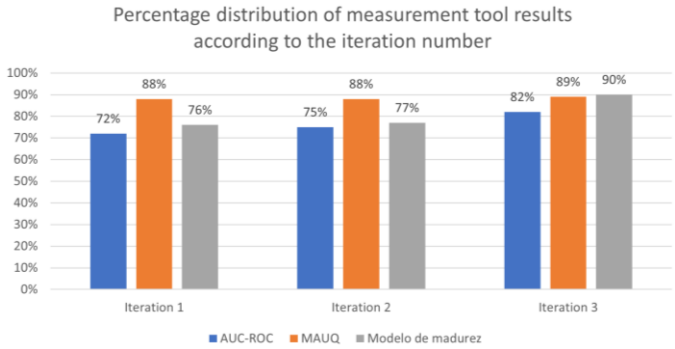

Fig. 4 Bar chart of the results of each measuring tool

Table 5 Calculations of the average additional accuracy of the maturity model

	Iteration 1	Iteration 2	Iteration 3
|Maturity Model - AUC-ROC|	4%	2%	8%
|Maturity model - MAUQ|	12%	11%	1%
Average of each iteration	8%	7%	5%
Average of the sum of iterations	6%		

Our proposal to put together these 2 approaches and our category of depression detection tool suitability has allowed us to have a more accurate assessment on the maturity of a depression detection chatbot, as we were able to obtain 6% more accuracy, on average, when evaluating these conversational agents.

V. Conclusions and Perspectives

The experimental procedure used in this study included the evaluation of a depression detection chatbot called "Sequitur", which received improvements in classification model training and user experience. This led to an increase in performance, depression detection tool suitability and usability, which resulted in a 24% increase in the quality score and a maturity level.

From the research conducted, we can also conclude that the combination of usability metrics, performance and the proposed suitability category allow us to measure the usability and effectiveness of chatbots as software products with greater accuracy, since, thanks to the fact that our maturity model considers the category of suitability, usability and performance in the evaluation variables, we have been able to achieve an increase in accuracy of 6% compared to other ways of evaluating this type of chatbots.

On the other hand, we believe that our experiments would have been more accurate if more artificial intelligence-based depression detection chatbots had been available, so there is a clear need for future research to improve and extend the scope of the proposed maturity model. These could determine the

addition of necessary approaches and improvements to the metrics already implemented in the model. Furthermore, we believe that our research can be complemented with a study that analyzes the difference in performance of classification models between chatbots trained with datasets based on depression detection tools and other information sources.

REFERENCES

[1] Alanzi, T. M. (2022). Users' satisfaction levels about mHealth applications in post-Covid-19 times in Saudi Arabia. PLoS ONE, 17(5 May). https://doi.org/10.1371/JOURNAL.PONE.0267002

[2] Angus, L., Goldman, R., & Mergenthaler, E. (2008). Introduction. One case, multiple measures: an intensive case-analytic approach to understanding client change processes in evidence-based, emotion-focused therapy of depression. Psychotherapy Research : Journal of the Society for Psychotherapy Research, 18(6), 629-633. https://doi.org/10.1080/10503300802430673

[3] Atoum, I. (2020). A novel framework for measuring software quality-in-use based on semantic similarity and sentiment analysis of software reviews. *Journal of King Saud University - Computer and Information Sciences*, *32*(1), 113-125. https://doi.org/10.1016/J.JKSUCI.2018.04.012

[4] Barletta, V. S., Caivano, D., Colizzi, L., Dimauro, G., & Piattini, M. (2023). Clinical-chatbot AHP evaluation based on "quality in use" of ISO/IEC 25010. International Journal of Medical Informatics, 170. https://doi.org/10.1016/J.IJMEDINF.2022.104951

[5] Braunschneider, L. E., Lehmann, M., Magaard, J. L., Seeralan, T., Marx, G., Eisele, M., Scherer, M., Löwe, B., & Kohlmann, S. (2021). GPs' views on the use of depression screening and GP-targeted feedback: a qualitative study. Quality of Life Research, 30(11), 3279-3286. https://doi.org/10.1007/S11136-020-02703-2

[6] Burchert, S., Kerber, A., Zimmermann, J., & Knaevelsrud, C. (2021). Screening accuracy of a 14-day smartphone ambulatory assessment of depression symptoms and mood dynamics in a general population sample: Comparison with the PHQ-9 depression screening. PloS One, 16(1). https://doi.org/10.1371/JOURNAL.PONE.0244955

[7] Caiado, R. G. G. G., Scavarda, L. F., Gavião, L. O., Ivson, P., Nascimento, D. L. D. M., & Garza-Reyes, J. A. (2021). A fuzzy rule-based industry 4.0 maturity model for operations and supply chain management. International Journal of Production Economics, 231. https://doi.org/10.1016/j.ijpe.2020.107883

[8] Chen, L., Bi, G., Yao, X., Tan, C., Su, J., Ng, N. P. H., Chew, Y., Liu, K., & Moon, S. K. (2023). Multisensor fusion-based digital twin for localized quality prediction in robotic laser-directed energy deposition. Robotics and Computer-Integrated Manufacturing, 84. https://doi.org/10.1016/j.rcim.2023.102581

[9] Gabrielli, S., Rizzi, S., Bassi, G., Carbone, S., Maimone, R., Marchesoni, M., & Forti, S. (2021). Engagement and effectiveness of a healthy-coping intervention via chatbot for university students during the COVID-19 pandemic: Mixed methods proof-of-concept study. JMIR MHealth and UHealth, 9(5). https://doi.org/10.2196/27965

[10] Garousi, V., & van Veenendaal, E. (2022). Test Maturity Model Integration: Trends of Worldwide Test Maturity and Certifications. IEEE Software, 39(2), 71-79. https://doi.org/10.1109/MS.2021.3061930

[11] Iivari, N., Sharma, S., & Ventä-Olkkonen, L. (2020). Digital transformation of everyday life - How COVID-19 pandemic transformed the basic education of the young generation and why information management research should care? International Journal of Information Management, 55. https://doi.org/10.1016/J.IJINFOMGT.2020.102183

[12] Lutkiewicz, K., Bieleninik, Ł., Cieślak, M., & Bidzan, M. (2020). Maternal-Infant Bonding and Its Relationships with Maternal Depressive Symptoms, Stress and Anxiety in the Early Postpartum Period in a Polish Sample. International Journal of Environmental Research and Public Health, 17(15), 1-12. https://doi.org/10.3390/IJERPH17155427

[13] Niedermaier, S., Zelenik, T., Heisse, S., & Wagner, S. (2022). Evaluate and control service and transaction dependability of complex IoT systems. Software Quality Journal, 30(2), 337-366. https://doi.org/10.1007/S11219-021-09556-Z

[14] Romero, M., Casadevante, C., & Montoro, H. (2020). How to create a psychologist-chatbot. Papeles Del Psicologo, 41(1), 27-34. https://doi.org/10.23923/PAP.PSICOL2020.2920

[15] Samaras, A. D., Moustakidis, S., Apostolopoulos, I. D., Papandrianos, N., & Papageorgiou, E. (2023). Classification models for assessing coronary artery disease instances using clinical and biometric data: an explainable man-in-the-loop approach. Scientific Reports, 13(1), 6668. https://doi.org/10.1038/S41598-023-33500-9

[16] Sanchez-Adame, L. M., Mendoza, S., Urquiza, J., Rodriguez, J., & Meneses-Viveros, A. (2021). Towards a Set of Heuristics for Evaluating Chatbots. IEEE Latin America Transactions, 19(12), 2037-2045. https://doi.org/10.1109/TLA.2021.9480145

[17] Saylor, C. F., Finch, A. J., Spirito, A., & Bennett, B. (1984). The children's depression inventory: a systematic evaluation of psychometric properties. Journal of Consulting and Clinical Psychology, 52(6), 955-967. https://doi.org/10.1037//0022-006X.52.6.955

[18] Souza-Pereira, L., Ouhbi, S., & Pombo, N. (2021). Quality-in-use characteristics for clinical decision support system assessment. *Computer Methods and Programs in Biomedicine*, *207*, 106169. https://doi.org/10.1016/J.CMPB.2021.106169

[19] Thompson, L. I., & Jones, R. N. (2020). Depression screening in cognitively normal older adults: Measurement bias according to subjective memory decline, brain amyloid burden, cognitive function, and sex. Alzheimer's & Dementia (Amsterdam, Netherlands), 12(1). https://doi.org/10.1002/DAD2.12107

[20] Toader, D. C., Boca, G., Toader, R., Măcelaru, M., Toader, C., Ighian, D., & Rădulescu, A. T. (2020). The effect of social presence and chatbot errors on trust. Sustainability (Switzerland), 12(1). https://doi.org/10.3390/SU12010256

[21] We are all agents of change in the fight against depression " UDEP Today (n.d.). Retrieved March 30, 2023, from https://www.udep.edu.pe/hoy/2023/01/todos-somos-agentes-de-cambio-en-la-lucha-contra-depresion/

[22] Unudulmaz, A., Cingiz, M. Ö., & Kalıpsız, O. (2022). Adaptation of the Four Levels of Test Maturity Model Integration with Agile and Risk-Based Test Techniques. *Electronics 2022, Vol. 11, Page 1985*, *11*(13), 1985. https://doi.org/10.3390/ELECTRONICS11131985

[23] Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape | " Chatbots " et agents conversationnels en santé mentale : une revue de l'environnement psychiatrique. Canadian Journal of Psychiatry, 64(7), 456-464. https://doi.org/10.1177/0706743719828977

[24] Zhou, L., Bao, J., Setiawan, I. M. A., Saptono, A., & Parmanto, B. (2019). The mhealth app usability questionnaire (MAUQ): development and validation study. JMIR MHealth and UHealth, 7(4). https://doi.org/10.2196/11500