Interaction Design and Architecture(s) Journal - IxD&A, N.59, 2023-24, pp. 79 - 94
DOI: 10.55612/s-5002-059-003

College students-in-the-loop for their mental health: a case of AI and humans working together to support well-being

Vanessa de Cássia Alves[1], Franco Eusébio Garcia[1], Conrado Saud[1]
Augusto Mendes[1], Helena Medeiros Caseli[1], Vivian Genaro Motti[2]
Luciano de Oliveira Neris[1], Tais Blecher[1] and Vânia P. Almeida Neris[1]

1 Federal University of Sao Carlos (UFSCar), Brazil
2 George Mason University (GMU), United States
vania.neris@ufscar.br

Abstract. Technology plays a relevant role in mental health. Specifically, integrating pervasive technologies with artificial intelligence (AI) holds promising potential to collect users' data, monitor individuals daily, and support treatment. However, the lack of trust in the collected data is a common limitation of prior work on mental health and technology. This paper proposes involving the user in a Human-in-the-loop approach as a solution to deal with the lack of accuracy of data collected through pervasive technology. In our study, end users judged and evaluated the data collected at two different times: before training the computational model, which would be later used for classification; and afterward to evaluate newly collected data that would be predicted and classified by the model. The solution proposed was implemented and tested in a project related to depression in college students. The results indicate positive reactions to the predicted classifications.

Keywords: Human-in-the-loop, artificial intelligence, mental health, depression, mobile sensors, wearable, digital phenotyping, college students.

1. Introduction

Mental health issues are a serious problem worldwide. In particular, depression, the mental health issue studied in our research study, causes significant distress as well as impairment in social, occupational, or other important areas of functioning in the individual's life. Major depression can be diagnosed when a person experiences five or more of the following symptoms for at least two weeks: decrease or increase in weight or appetite; insomnia or hypersomnia; agitation or psychomotor retardation; fatigue or loss of energy; feelings of worthlessness or guilt; decreased concentration, or indecision and/or recurring thoughts of death (DSM-V). To conclude the diagnosis at least one symptom reported must be a depressed mood or loss of interest or pleasure. Data before the COVID-19 pandemic indicated that 7 to 26% of the American population suffered from depression [1].

In the university environment, the context chosen for study in this project, the prevalence of depression appears to be higher than in the rest of the population.

Lauckner et al. [2] state that “college students bear a disproportionate burden of depression when compared to the general population”. Ibrahim, et al. [3] carried out a systematic review of studies on the prevalence of depression in college students around the world and found that the average prevalence of depressive symptoms was 30%. similar study using the Beck Depression Inventory scale also found a 30.6% prevalence of depressive symptoms among Brazilian medical students [4].

Mental health treatments require time and monitoring by a specialized professional. Moreover, there is a consensus that early identification helps the treatment of mental health problems. In general, the identification of the problem does not occur punctually, through a single episode of discomfort, but requires observation of the individual's behavior over time, including their physical and psychological reactions [5]. In this context, computing plays a relevant role. Specifically, the usage of pervasive technologies integrated with artificial intelligence (AI) enables individuals in daily actions, data collection as well as the identification of mental health symptoms such as lethargy or agitation, fatigue, and poor sleep quality. Lastly, technology might even automatically process self-report data.

Melcher et al. [6] conducted a clinical review of 25 mental health digital phenotyping studies that collected data from college students. The studies lasted an average of 42 days and had an average enrollment of 81 participants. The most common data came from location sensors, as well as accelerometers and social information. These data were used as a proxy to indicate individual behaviors such as sleep, exercise, and social interactions. Other data included mood, anxiety, and stress. The authors concluded that, between the studies, there is still a lot of heterogeneity in the collection and analysis methods.

Trust in the data collected in mental health studies is a common limitation of the studies analyzed. Open questions remain such as: Does the data collected actually represent what was faced by the student? Does a low sleep score suggest insomnia or is it the result of a night after a great party? To address the limited accuracy of data collected through pervasive technology this paper proposes involving the user in a Human-in-the-loop (HITL) approach. To investigate this approach, we conducted a user study, in which end users judged and evaluated the data collected twice. First, before training the computational model (to be further used for classification); and afterward to evaluate newly collected data (to be classified by the model). The solution proposed was implemented and tested in a larger project aimed at studying technology and depression in higher education. The results suggest a positive reaction to the predicted classifications with more than 70% enjoying the use of the recommendations.

This paper is organized as follows: section 2 summarizes the background describing other projects that collect data from college students to support their mental health and introduces the theory that inspired the proposed approach of involving humans in the process of monitoring and intervening in AI predictions. Section 3 presents the Amive project led by the authors, briefly describing two main phases: the construction of the AI model to recommend conversation themes by a chatbot; and the classification of several data by this model when in use by college students. Section 4 presents the project HITL solution proposed to address the lack of data accuracy. Section 5 discusses the use of the solution proposed by college students, and Section 6 concludes the paper.

2. Background

2.1 AI for college students' mental health

The review by [6] discusses projects carried out during the period ranging from 2014 to 2020. Each project analyzed had a different focus. For this paper, the studies by [7] and [8] are of particular interest. These studies were chosen because they are closer to the Amive project in their goal and because of the data collected.

In [7], 48 students from Dartmouth College participated in a research lasting for a total of 10 weeks. The study collected Ecological Momentary Assessments (EMAs) answered by participants through surveys made available throughout each day. According to [9], EMA uses repeated real-time data collection regarding college students' behaviors and experiences in their natural environments. On average, 8 EMAs were triggered per day for each student. The types of questionnaires employed in this study were the Patient Health Questionnaire (PHQ-9), to measure the level of depression, the Perceived Stress Scale (PSS), to measure the level of stress, the Flourishing scale, to measure the level of psychological well-being, UCLA loneliness scale, to measure the level of loneliness, and Big five inventory (BFI), to measure personality traits. In addition to these data, the following behavioral data were also collected using the BeWell and StudentLife applications: activity (standing, walking, running, cycling, driving), sleep duration, and sociability (number of independent conversations and their duration). The StudentLife app also collects accelerometer, proximity, audio, light sensor readings, location, co-location, and app usage. Through the data collected in their study, researchers were able to identify correlations between data from smartphone sensors and students' responses to the well-being questionnaires.

Boukhechba et al. [8] also carried out a study focusing on mental health. The study lasted two weeks and involved 72 students from the University of Virginia. In the first stage of the study, the Social Interaction Anxiety Scale (SIAS) tests were applied to measure suffering in social situations. The Depression, Anxiety, and Stress Scales (DASS) were used to assess symptoms of depression; whereas the Positive and Negative Affect Schedule (PANAS) assessed general positive and negative mood. During two weeks, participants responded to seven EMA questionnaires per day, through the Sensus application installed on each student's smartphone. Furthermore, in this same application, it was possible to retrieve information such as GPS location data every 150 seconds, accelerometer data (1 HZ), and call-end text logs over the study passively. With this study, researchers were able to compare data collected through daily questionnaires, passive data, and the results obtained in mental health tests.

Studies [7] and [8] carried out comparisons between students' behavior and mental health and also tried to associate these results with the academic performance of each student. Besides collecting data from sensors, both projects rely on several EMAs per week, which may overload the students with too many questionnaires.

Other studies focusing on students' mental health and investigating the use of AI are cited below. Ware et al. [10] made use of Support Vector Machine (SVM), and collected Global Positioning System (GPS) and Wi-Fi information from mobile

devices to collect participants' location information, such as places visited, time spent in movement and distance covered, to identify signs of agitation or retardation, fatigue and changes in sleep. Farhan et al. [11] used SVM to evaluate whether behavioral features extracted from students' smartphones, including GPS data, could monitor and predict depression. The Random Forest algorithm is explored by Narziev et al. [12] to identify signs of depressive mood, loss of interest or pleasure in activities, and other symptoms related to depression in students. Kim et al. [13] investigated SVM, Random Forest, and Extreme Gradient Boosting (XGBoost) jointly to create classifiers for depression prevention using sleep features that are extracted from smartphones.

Other works conducted studies with college students applying algorithms in a specific manner, covering techniques such as Singular Spectrum Analysis (SSA) and Leave-one-out cross-validation (LOOCV). A study conducted by [14] used SSA to predict depression by analyzing participants' heart rate, physical activity, and sleep monitoring data. Masud et al. [15] used LOOCV with regression models to detect agitation or fatigue with accelerometer and gyroscope data.

Unlike prior work, the Amive project deals with the lack of accuracy problem giving the humans the primacy to evaluate the data collected before they are processed.

2.2 HITL approaches

The Amive project aims to detect a Possible Depressive Profile (PDP) in college students. After detecting a PDP student based on the classification of their symptoms, the project provides intervention content to the student. To carry out a classification corresponding to the reality of each student, it is important to check if the data collected is accurate, as there may be disagreement between the values obtained by mobile sensors and the reality experienced by the student.

To address this need, we adopted a Human-In-The-Loop (HITL) approach. HITL indicates the presence of human interaction with the system [16], which can be carried out through dialogues or notes made by the individual in the system, to improve the quality of the Machine Learning (ML) models developed in the system. This HITL approach is aligned with the concept of Human-Centered Artificial Intelligence (HCAI), described by [17] as a two-dimensional framework composed of human control and computer automation. In HCAI, the human being and the computer can have different degrees of control over the situation and the system, it is up to the designer to decide on the level of automation necessary to control the technology.

Within this HCAI concept, the presence of Usable AI and Useful AI elements can also be noted in the project, which, are respectively defined as a system that is easy to use by its end users and must offer adequate and reliable results [18]. These elements are particularly important because the solution proposed is aimed at helping people with PDP. This project also makes use of Explainable Artificial Intelligence (XAI) concepts [19], which aim to make the system's behavior more intelligible to humans.

In ML, models with human interaction in the learning loop can be classified as Interactive Machine Learning (IML). IML is characterized by containing the model,

users, data, interface, and ML model as fundamental parts, as stated in [20]. The user can be present at any time during the learning process from beginning to end, collaborating with validation, data cleaning, and correction of results [21]. The user interface is crucial for communication between the user and the system. In IML, concepts of Human-Computer Interaction (HCI) are also inserted thanks to the use of the user interface. The solution we present in Section 4 follows these concepts.

3. Amive project

Amive is an acronym in Portuguese for three words: Friend, someone with whom you can talk, who cares and aims to support your well-being; Virtual, as it is a computational solution delivered remotely; and Specialized, as it is based on content endorsed by mental healthcare experts. The project aimed to build a computational infrastructure for autonomous, and real-time identification and intervention with PDP users.

Due to the requirement of predicting a PDP among users based on the collected data, the implementation of Amive required end user participation. An initial application (see Fig. 1) has been implemented to collect data from college students. These data were used to train computational models capable of identifying behavioral signs and symptoms of interest.

3.1. Model training

In this initial phase, data from social networks and mobile sensors was collected from 89 college students at the Federal University of Sao Carlos (UFSCar) for five weeks. These students included people with and without depression, as measured by the results of PHQ-9. Fig. 1 summarizes the frequency of data collection, data sources, and data collected to build the computational model for the Amive project. Most data was either submitted or provided to the application by the participants themselves. This was the case for sensor data (heart rate, exercises, sleep, and steps), collected by smartwatches and sent to the application by the participants; answers to questionnaires: PHQ-9, WHO Disability Assessment Schedule (WHODAS 2.0), and a self-report survey to ask how the user had been feeling at the time of use; and Facebook posts. As Twitter provided an Application Programming Interface (API), a back-end system collected tweets from participants who authorized it.

As the collected data had different time measures, it was unsuitable to train traditional ML models without preparation. For instance, sensor data often was provided as a time series, while questionnaires had discrete submission times. Thus, to build training data for supervised learning, data have been grouped in daily interval windows. This required aggregating data based on start and end timestamps, as well as defining valid ranges for the PDP classification used as the label (expected classification for whether a participant was depressive at the time).

The aggregation resulted in a dataset defined as follows. Each row provided data for a given day after the start of the data collection. Columns of interest included the predominant self-reported subjective feeling; the average heart rate; average wake

heart rate; heart rate variance; time that the student tried to sleep; quality of the sleep as provided by the smartwatch; time that the student woke up; quality of naps as provided by the smartwatch; number of steps; walked/run distance; whether the student exercised; and the exercise that the student performed for the most time. The expected outcome was provided by the sum of PHQ-9 points for the fortnight. Participants were labeled as PDP if the sum was greater or equal to 9 (the maximum possible score is 27). For aggregation and imputation, the authors implemented data analysis scripts using the Python and R programming languages, with libraries including Pandas and Mice. The data has been aggregated using Python and then imputed using R.

The resulting dataset allowed training supervised learning models, including Logistic Regression, Decision Trees, Support Vector Machines; Naive Bayes; K-Nearest Neighbors (KNN); XGBoost, and Random Forest. Thus, the trained models could try to predict whether a given participant was PDP on a given day based on the same input data (that is, collected and processed in the very same way). The models were implemented in Python using Scikit-learn and statsmodels. They were trained using a split of 70% of the data for training, and a 30% split for testing, exploring different combinations of features to try to find a model that could provide better predictions overall. To serve as a baseline, initial attempts used single features from an individual source (for instance, heart rate, sleep data, or self-reported feelings). The accuracy of these classifiers varied among models; the best classifiers predicted correctly between 53% to 71% of the test split. The self-reported feeling provided the best results, from the decision trees, support vector machines, and logistic regression models.

Afterward, the goal was to verify whether a multifactorial model (*i.e.*, combining data from two or all data sources) could provide better classification than a single factorial one. To test this hypothesis, several combinations of subsets of features from the data sources (sensors, questionnaires, and self-reported feeling) have been generated from a script that combined/dropped features based on the total number of missing values in the resulting dataset. The lengths and number of users of these datasets varied according to the availability of the data provided per day and per source by the participants.

Three of these combinations have been chosen as new datasets to train and test the original models once again. This time, five holdouts splits have been applied to each model. As the accuracies of the predictions have varied among each holdout split, the next sentences describe the result from the split with the highest accuracy. The first dataset combined sensor data with self-reported emotions. The resulting aggregated dataset had 373 rows with data from 32 participants (with roughly 45% users from the PDP-group, and 55% from the non-PDP group); in this case, logistic regression with Ridge regularization (alpha 1.0) provided the best result, correctly predicting 81% of the test set. A second dataset has been generated as a variation of the first, with 294 rows from 25 participants (roughly 46% from the PDP-group, and 54% from the non-PDP-group). This time, logistic regression with LASSO regularization (alpha 100.0) and XGBoost predicted 78% of the test set correctly. Finally, the third dataset had features from all data sources, with 72 rows from 10 participants. However, in addition to the lower sample size, this last dataset was unbalanced, as almost 82% of the labels were from the non-PDP group. Considering these caveats, logistic

regression with Ridge regularization (with alpha 0.01 and 1.0) provided the best results for this third dataset, predicting 95% of the test set correctly. The following study used the trained models with data from new participants. This time, the logistic regression models (that performed well with the original test data) provided poor predictions for the new participants. The accuracy of predictions from all models was significantly lower, usually predicting between 30% to 50% of the new inputs correctly. The decision trees model provided the best results overall, with only 58% accuracy.

In parallel, computational models were trained to classify 18 signs of interest (depressive symptoms and risk or protective factors) from a dataset manually annotated by mental health specialists containing 780 posts, resulting in 2304 labeled spans of text indicative of one or more of the evaluated signs. Again, several supervised learning algorithms (the ones mentioned above) and also fine-tuning of pre-trained BERT models were used. A multi-task architecture was also evaluated in the case of the pre-trained models since the regularizing effect of joint learning could mitigate overfitting (a concern given the small sample size of some signs, which could be as low as 15 positive instances). The GoEmotions dataset was selected for the auxiliary task of emotion classification, given the similarity to the sign classification task in terms of data sourcing (social media texts), subject matter (understanding a patient's emotional state plays a large role in evaluating their condition) and task structure (fine-grained multilabel classification). The multi-task model architecture consisted of the pre-trained models for the shared parameters and a linear classification head for the task-specific parameters. These computational models obtained average precision scores (that is, the area under the precision-recall curve) ranging from 28-86% according to the evaluated sign. Overall, the multi-task fine-tuned neural models were the ones with the best results and were thus selected for usage. Examples of symptoms include sadness or depressive humor; tiredness; suggestions of suicide; hopelessness; fear and anxiety; aggressivity; other risk factors; and healthcare and well-being. The textual data was transformed into features (e.g. TF-IDF, embeddings, POS-tags) for feature-based models and contextual embeddings for fine-tuning the pre-trained language models.

Amive also included a conversational agent (chatbot). The chatbot was designed following the guidelines presented in [23]. The chatbot could dialogue with students following predefined scripts written by the healthcare professionals, part of the research team. Dialogues conveyed relevant health education, care, and prevention content to the participants. The healthcare professionals could tag each dialogue based on the symptoms that it tried to address. These symptoms matched the ones predicted by the previously described neural network trained to evaluate text. This enabled dialogue suggestions using recommender systems. A simple recommendation system has been implemented to suggest dialogues based on the classification of text content posted by users. In this recommendation system, a dialogue recommendation resulted from a non-empty intersection between the tagged symptoms of the dialogue and the symptoms classified from a user's text messages. In other words, a dialogue was recommended if its tags included at least one of the symptoms identified from any of the user's text messages; hence the simplicity of the recommendation system. This minimized the chance of a potentially useful dialogue being excluded from recommendations. Overall, the healthcare team considered it was better to suggest a

dialogue that could be useful to a user – even if she/he ignored it –, than not suggesting it, and potentially missing an opportunity to aid the user. The chatbot and recommendation system were implemented as a mobile application written in Dart, using the Flutter framework.

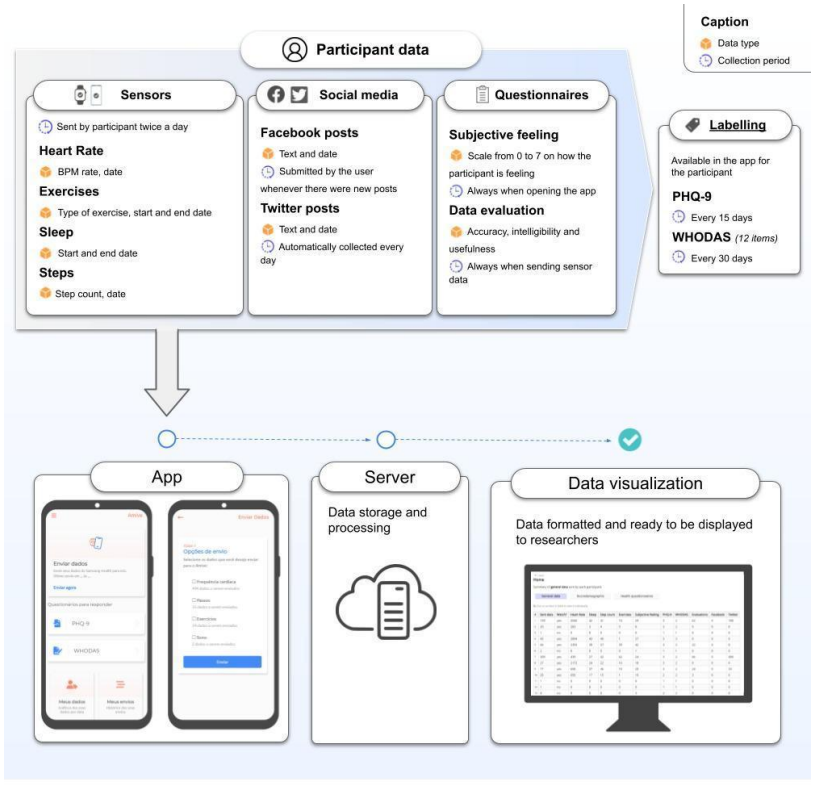

Fig. 1. Top: Data collected to build the computational model for the Amive project. Bottom: Modules implemented for data collection (App), storage and processing (Server), and visualization.

3.2. Model usage

Therefore, the complete Amive solution consists of an application that (i) collects data provided by the user, from a smartwatch, questionnaires, and social network posts; (ii) sends this data to a server where the database and computational models are deployed and running; (iii) receives information about identified symptoms and guidance for dialogue by classification models; (iv) allows the visualization of graphs on the data collected and dialogues with the student through the chatbot, as well as rating this data for accuracy, understandability, and utility; (v) provides a diary to

enable further extraction of possible symptoms from written contents; among other features. The data collected by the Amive application includes (i) heart rate, sleep quality, number of steps, and physical activity from mobile sensors present in a smartwatch; (ii) public texts posted by the student on the social networks Facebook and Twitter; (iii) list of answers about how the student was feeling at the time of the answer; and (iv) text produced by the student including daily life reports, through the diary functionality. Some of this data is aggregated and imputed by the back-end, evaluated by the trained AI algorithms, and sent back to the applications, which, then, can suggest content or actions for the user. The back end is also able to send e-mails to selected participants and healthcare professionals whenever a possible severe case of PDP is identified. Fig 2 represents the data and this general data flow for the Amive project.

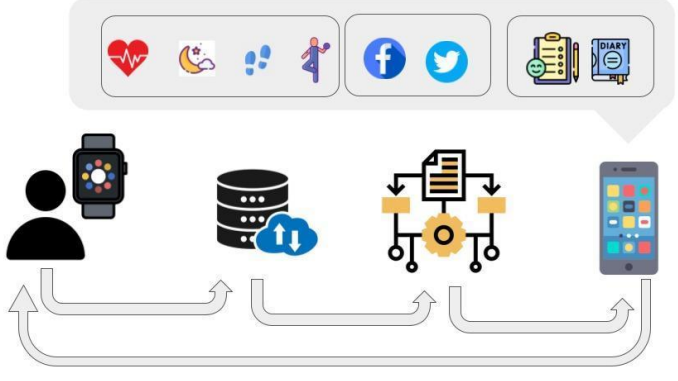

Fig. 2. Data and general data flow for the Amive project. Top: types of data collected. Bottom: The user collects data using a mobile phone and a smartwatch. Data is sent to the server using the app. Data is processed and classified in the server and dialogues/conversations are offered in the app.

4. Amive HITL solution

In this project, users are undergraduate and graduate students. The data comes from the students themselves. The user interface (UI) is composed of navigation menus to access the functionalities of the mobile application, a natural language interface of a personal diary, questionnaires, and a chatbot, with previously created dialogues, without generating dialogues at run time, just with the selection of options. As explained in section 3.1, the dialogues created for the project were based on determinants of suffering raised by psychologists with experience in mental health for college students.

For training the model (section 3.1), the data collected from each student underwent validation, carried out by the students themselves. When validated, approved data was used for training; otherwise, it was rejected and disregarded from

DOI: 10.55612/s-5002-059-003

further processing. At this stage, the user's presence in the system is seen in data validation, before the model training, as characterized by IML. In the model usage stage (section 3.2), the user is part of the loop and is responsible for validating the collected data, through the application interface. This validation helps the system decide whether or not to use the data for classification if this data is according to what was experienced by the user. In this sense, it is also possible to identify characteristics of a model that follows the HCAI guidelines, with the concepts of Usable AI and Useful AI (section 2.2), mainly because it is a solution aimed at helping people with PDP.

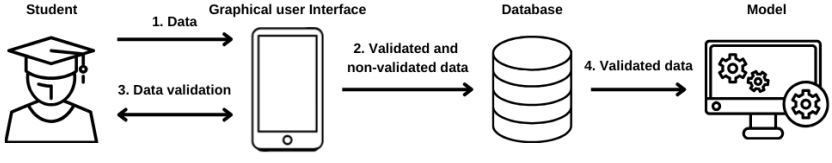

Fig. 3. HITL in the model training stage.

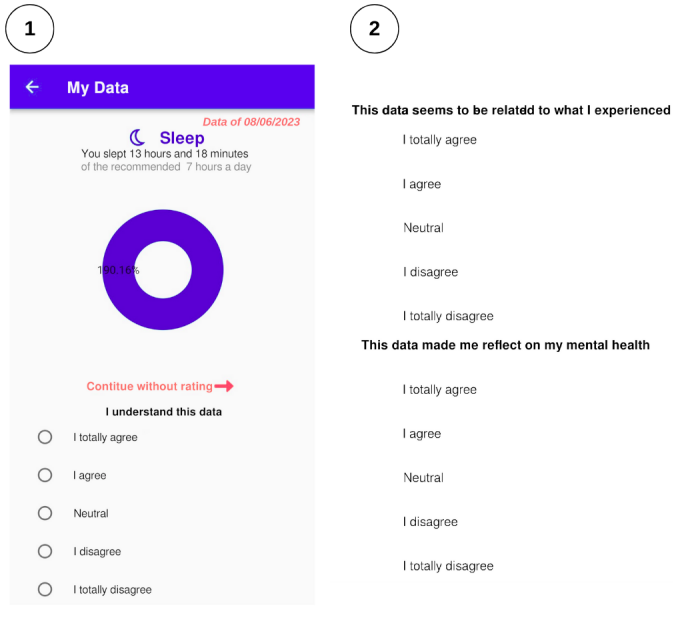

Fig. 4. Amive app user interface for HITL.

The diagram in Fig. 3 represents the cycle of user interaction with the system in the model training phase, which was carried out through the UI available in the Amive application. As illustrated in Fig. 3, training occurs as follows: (1) the user sends their data through the application, this data comes from sensors, diary texts, present in the application, tweets, and Facebook messages; (2) then validated and non-validated data is persisted into the database; (3) she/he validates the data obtained by the sensors

according to a questionnaire containing the following three questions: (a) “I understand this data”, (b) “This data seems to be related to what I experienced”, (c) “This data made me reflect on my mental health”, which were answered using a Likert Scale with the following answers “Totally agree”, “Agree”, “Neutral”, “Disagree”, “Totally disagree”. Fig. 4 shows the user interface with the three questions. This validation process is requested for the student’s data collected by the smartwatch: sleep, physical exercise, heart rate, and the number of steps. Sentence (b) was used for the HITL data validation; (4) only user-validated sensor data is used to train the ML model, as well as for predictions of the trained models.

The diagram in Fig. 5 represents the cycle of user interaction with the system through the UI, in the stage of using the trained model. In this diagram it is possible to see the order in which the interaction between the user and the system occurs: for the data to be processed by the model to be classified, first (1) the data from the sensors, the diary texts, present in the application, tweets and Facebook messages are collected; and (2) all data, whether validated as correct or not, is stored in the database; (3) validation of sensor input data is carried out by the user through the interface, using the same three questions and the same Likert Scale as in the model training stage. At this stage, the student confirms or rejects the correctness of the information collected through the sensors of their own devices; (4) only the data validated as correct goes to (5) classification in the ML models, at this stage, the sensor data is forwarded to the ML model and the text data is sent to a text classifier developed within the scope from the project; (6) based on the classification of PPD or Non-PDP (true/false), obtained by the ML model from sensor data, the invitation to chat is made or not. Based on the textual classifications (symptom identification), obtained by the Natural Language Processing model, the application can select the dialogues, (7) to be presented to the user classified as PDP. If no symptoms have been detected, no particular conversation is suggested and the following message is presented to the user: "No specific recommendations for you. Write some entries in your journal. If you choose to share Twitter messages or send Facebook messages, they will also be used for recommendations. Since there are no suggestions at this time, would you like to see a list of all available dialogues?". Even without particular recommendations, the user can start one of the available dialogues or, if she/he prefers, she/he can choose, from the dialogue menu, a dialogue that is convenient for her/him.

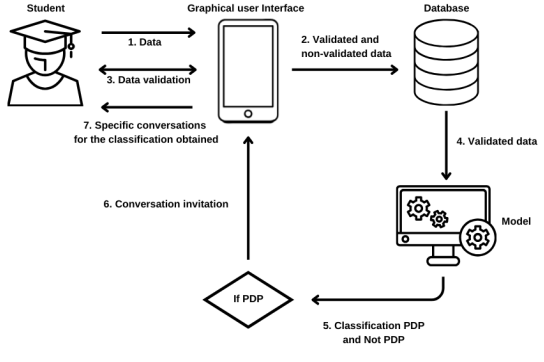

Fig. 5. HITL in the solution use stage.

5. Evaluation and discussion

Following the HITL model, to evaluate the usefulness of conversations, students rate the dialogues through the user interface in the chatbot. This evaluation consists of two questions asked at the end of each dialogue: “How do you evaluate the experience of chatting with Amive?”, with the possible answers: “I liked talking to Amive”, “I am neutral” and “I didn’t like talk to Amive”; “How do you evaluate the impact of the conversation?”, with the possible answers: “It helped me”, “I'm neutral”, “It didn't help me”, as shown in Fig. 6. The answers given to these two questions allow the evaluation of the delivery of the intervention made through conversations, checking whether it supported the participants’ well-being.

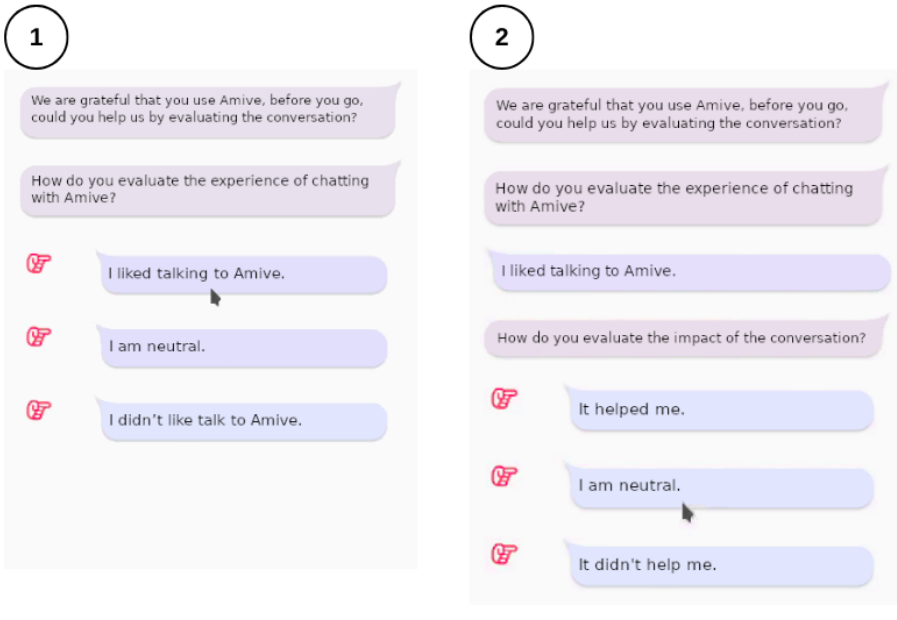

Fig. 6. Evaluation questions at the end of dialogues.

For the clinical study of the project, a new call for participation was made for student volunteers to register. At the time of enrollment, phase 1 of the study, these students responded to a questionnaire with sociodemographic characteristics and a self-administered scale questionnaire PHQ-9, validated for the Portuguese language [22], consisting of nine questions. Students who presented symptoms compatible with mild depression (score between 5 and 9 on the PHQ-9, who did not actively think about hurting themselves) or moderate depression (score between 10 and 14 on the PHQ-9, who also did not consider hurting themselves) were invited to join the phase 2 of the study, called monitoring.

Following the IRB-approved protocol of the study, those with moderately severe symptoms (sum between 15 and 19 points on the scale), severe (sum between 20 and 27), or active suicidal ideation (response greater than 0 in item 9 of the PHQ-9) were not recruited for phase 2 (monitoring). They were instructed to seek help and were offered reception and listening by a health professional, with the intention of psychoeducation and referral to a health service. These participants also received the results of the scale scores and were guided, through automatic messages, to contact a health service to begin health monitoring.

Once this selection was made, 20 students actively participated in using the complete system built, sending data for three weeks, and interacting with the developed application. Of these 20 students, 18 answers were given to the questions at the end of each dialogue. From 18 answers given to the questions that evaluate the dialogues, it was possible to analyze the impact and satisfaction of these dialogues in helping study participants raise awareness about the symptoms of depression. Fig. 7 and Fig. 8 show the number of positive, negative, and neutral responses obtained for each question, Fig. 7 presents the answers to the first question, and Fig. 8, to the second question.

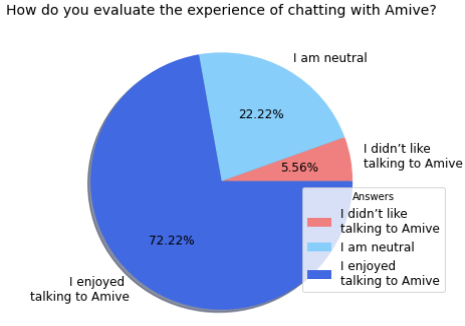

Fig. 7. Percentages on how users rated the experience of chatting with Amive.

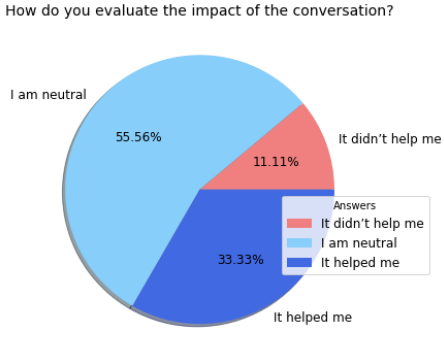

Fig. 8. Percentages on how users rated the impact of the conversations.

Although most students (13 of the 18 answers) enjoyed talking to Amive, only six participants considered that the dialogue helped in some way. On the other hand, only one participant did not like talking to Amive and two felt that the dialogue did not help them. Additionally, 10 participants were neutral about the impact of the dialogue and four students were neutral about the enjoyment of talking to Amive.

The results obtained suggest that it may be necessary to richer content and/or restructure existing dialogues, aligned with the standard set of the existing dialogues as they were well accepted by the students in general. Due to the way these dialogues were constructed, following a decision tree model, a response in a branch could potentially eliminate a section of greater interest to the person. Regardless of this limitation, the students considered the adopted conversational user interface approach an interesting solution.

The correct use of the HITL approach designed for Amive requires user validation. To fulfill this requirement, the user must verify the data immediately after sending this data. Otherwise, the system may classify undesired data. Therefore, a current limitation of this work relies on the timings for user validation. Although users can validate their data at any time, the classification is performed automatically by the back-end of Amive at fixed hourly times. As a consequence, there is no guarantee that unvalidated data will not be processed (if not rated by the user before use). To deal with this, possible solutions are to engage users in classifying more data immediately after collecting or even investigating new ways of predicting data accuracy.

6. Conclusion and future work

This paper presented a possible solution to the problem of lack of accuracy of data to be processed in AI classification models. Amive, the solution described in this paper, delivers specialized mental health content in the format of dialogues through a chatbot. The content generated is based on symptoms identified from data collected through sensors and texts. The solution considers the involvement of college students in judging the accuracy of data via the UI available in an application immediately after sending the data. In the application, end-users were able to evaluate their data as a means to improve the relevance of the classifications. If the users rated the data as inaccurate, the evaluated data was not used to train classification algorithms at the first moment, nor to predict their depressive status. The solution was evaluated by 18 students who reported that they enjoyed talking to the chatbot.

The project presented is an example of AI supporting everyday actions; in this case, delivering mental health education content to college students with depression. For such solutions, the active involvement of users in the design and evaluation process of the solution is as important as the efforts to build the systems. The presented HITL approach illustrates a possibility of joint work in which end users have a voice in the use of artificial technologies by acting using the provided UI, instead of being mere data providers. Future works include improving dialogues and investigating personalization, further studies on better-associating data from mobile sensors and depressive symptoms, and new ways of improving users' participation in judging data while talking to the chatbot.

Acknowledgments. The authors thank Prof. Hélio Crestana Guardia and Amanda Basso de Oliveira for their previous contributions to the Amive system, to Prof. Jair Barbosa Neto, Vinicius Fratta Fritz and Lucas do Carmo Lima for their clinical support to students, and to Prof. Larissa Martini Barbosa, Prof. Heloisa Figueiredo Frizzo, Rafael dos Santos Elias and Daniela Gonzaga Fernandes for their contributions in the design of the chatbot. This research was funded by the São Paulo Research Foundation (FAPESP) - grants 2020/05157-9, 2022/16364-0, 2022/09173-4.

Vanessa Alves: Investigation, Formal Analysis, Writing - review and editing. **Franco Garcia**: Software, Data Curation, Writing - review and editing. **Conrado Saud**: Software, Data Curation, Validation, Writing – review and editing. **Augusto Mendes**: Methodology, Software, Data Curation, Writing – review and editing. **Helena Caseli**: Conceptualization, Methodology, Writing – review and editing, Supervision, Resources. **Vivian Motti**: Conceptualization, Methodology, Writing – review and editing. **Luciano Neris**: Conceptualization, Methodology, Writing – review and editing, Resources. **Tais Bleicher**: Conceptualization, Methodology, Writing – review. **Vania Neris**: Conceptualization, Methodology, Writing – review and editing, Supervision, Funding acquisition.

References

1. Eichstaedt, J. C.; Smith, R. J.; Merchant, R. M.; Ungar, L. H.; Crutchley, P.; Preoţiuc-Pietro, D.; ASCH, D. A.; Schwartz, H. A. Facebook language predicts depression in medical records. PNAS, v. 115, n. 44, October, 2018. DOI: 10.1073/pnas.1802331115.
2. Lauckner, C.; Hill, M.; Ingram, L. A. An exploratory study of the relationship between social technology use and depression among college students. Journal of College Student Psychotherapy, 34(1), 33-39, 2020. DOI: https://doi.org/10.1080/87568225.2018.1508396.
3. Ibrahim, A. K.; Kelly, S. J.; Adams, C. E.; Glazebrook, C. A systematic review of studies of depression prevalence in university students. Journal of Psychiatric Research. Volume 47, Issue 3, Pages 391-400, 2013. DOI:10.1016/j.jpsychires.2012.11.015.
4. Pacheco, J. P., Giacomin, H T., Tam, W. W., Ribeiro, T. B., Arab, C., Bezerra, I. M., & Pinasco, G. C. Mental health problems among medical students in Brazil: a systematic review and meta-analysis. Brazilian Journal of Psychiatry, 39(4), 369-378. 2017. DOI: 10.1590/1516-4446-2017-2223.
5. APA. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-V). Arlington, VA: American Psychiatric Association, 2013.
6. Melcher J, Hays R, Torous J. Digital phenotyping for mental health of college students: a clinical review. Evid Based Ment Health. 2020 Nov;23(4):161-166. Epub 2020 Sep 30. PMID: 32998937; PMCID: PMC10231503. DOI: 10.1136/ebmental-2020-300180.
7. Wang R, Chen F, Chen Z, et al. StudentLife: Using smartphones to assess mental health and academic performance of college students. Mobile Health - Sensors, Analytic Methods, and Applications, 2017. DOI: 10.1007/978-3-319-51394-2_2.

8. Boukhechba M, Daros AR, Fua K, et al. DemonicSalmon: monitoring mental health and social interactions of college students using smartphones. Smart Health 2018;9-10:192–203. DOI: 10.1016/j.smhl.2018.07.005.
9. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol. 2008;4:1-32. PMID: 18509902. DOI: 10.1146/annurev.clinpsy.3.022806.091415.
10. WARE, S. et al. Predicting depressive symptoms using smartphone data. Smart Health, Elsevier BV, v. 15, p. 100093, mar. 2020. DOI: https://doi.org/10.1016/j.smhl.2019.100093
11. FARHAN, A. A. et al. Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data. p. 1–8, 2016. DOI: 10.1109/WH.2016.7764553.
12. NARZIEV, N. et al. STDD: Short-term depression detection with passive sensing. Sensors, MDPI AG, v. 20, n. 5, p. 1396, mar. 2020. DOI: 0.3390/s20051396.
13. KIM, J.; HONG, J.; CHOI, Y. Automatic depression prediction using screen lock/unlock data on the smartphone. In: 2021 18th International Conference on Ubiquitous Robots (UR). IEEE, 2021. DOI: 10.1109/UR52253.2021.9494673.
14. DAI, R. et al. Multi-task learning for randomized controlled trials: A case study on predicting depression with wearable data. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., Association for Computing Machinery, New York, NY, USA, v. 6, n. 2, jul 2022. DOI:10.1145/3534591.
15. MASUD, M. T. et al. Unobtrusive monitoring of behavior and movement patterns to detect clinical depression severity level via smartphone. Journal of Biomedical Informatics, Elsevier BV, v. 103, p. 103371, mar. 2020. DOI: 10.1016/j.jbi.2019.103371.
16. Tomaszewski John E. Chapter 11 - Overview of the role of artificial intelligence in pathology: the computer as a pathology digital assistant. Stanley Cohen, Artificial Intelligence and Deep Learning in Pathology, Elsevier, 2021. Pages 237-262. DOI: 10.1016/B978-0-323-67538-3.00011-7.
17. Shneiderman B. Human-centered artificial intelligence: Reliable, safe & trustworthy. International Journal of Human–Computer Interaction, 36(6):495–504, 2020. DOI: 10.1080/10447318.2020.1741118.
18. Wei Xu. Toward human-centered AI: a perspective from human-computer interaction. Interactions 26, 4, 2019. Pages 42–46. DOI: 10.1145/3328485.
19. Gunning D, Stefik M., Choi J., Miller T., Stumpf S., and Yang G.-Z.. Xai-explainable artificial intelligence. Science Robotics, 4(37), Science Robotics, 4 (37) 18 December 2019, DOI: 10.1126/scirobotics.aay7120.
20. Chris J. Michael, Dina Acklin, Jaelle Scheuerman. On Interactive Machine Learning and the Potential of Cognitive Feedback. March 2020. DOI: https://doi.org/10.48550/arXiv.2003.10365.
21. Mosqueira-Rey E, Hernández-Pereira E, Alonso-Ríos D, Bobes-Bascarán J, Fernández-Leal A. Human-in-the-loop machine learning: a state of the art. Artificial Intelligence Review, 56, 08 2022. DOI: 10.1007/s10462-022-10246-w.
22. Santos, Iná S. et al. Sensibilidade e especificidade do Patient Health Questionnaire-9 (PHQ-9) entre adultos da população geral. Cadernos de Saúde Pública [online]. 2013, v. 29, n. 8, pp. 1533-1543. ISSN 1678-4464. DOI: 10.1590/0102-311X00144612..
23. Paula Maia de Souza, Isabella da Costa Pires, Vivian Genaro Motti, Helena Medeiros Caseli, Jair Barbosa Neto, Larissa C Martini, and Vânia Paula de Almeida Neris. 2022. Design recommendations for chatbots to support people with depression. In Proceedings of the 21st Brazilian Symposium on Human Factors in Computing Systems (IHC '22). Association for Computing Machinery, New York, NY, USA, Article 12, 1–11. DOI: 10.1145/3554364.3559119.