Predicting User Engagement in Longitudinal Interventions with Virtual Agents

Ha Trinh, Ameneh Shamekhi, Everlyne Kimani, Timothy W. Bickmore
Northeastern University
Boston, MA, USA
hatrinh, ameneh, kimani15, bickmore@ccs.neu.edu

ABSTRACT

Longitudinal agent-based interventions only work if people continue using them on a regular basis, thus identifying users who are at risk of disengaging from these applications is important for retention and efficacy. We develop machine learning models that predict long-term user engagement in three longitudinal virtual agent-based health interventions. We achieve accuracies of 74% to 90% in predicting user dropout in a given prediction period of the intervention based on the user's past interactions with the agent. Our models contain features related to session frequency and duration, health behavior, and user-agent dialogue content. We find that the features most predictive of dropout include number of user utterances, percent of user utterances that are questions, and the percent of user health behavior goals met during the observation period. Ramifications for the design of virtual agents for longitudinal applications are discussed.

CCS CONCEPTS

• **Human-Centered Computing** → Human Computer Interaction (HCI)

KEYWORDS

Engagement prediction, dropout prediction, longitudinal interventions, health interventions, conversational agents

ACM Reference format:

H. Trinh, A. Shamekhi, E. Kimani, and T. Bickmore. 2018. Predicting User Engagement in Longitudinal Interventions. In *Proceedings of the 18th ACM International Conference on Intelligent Virtual Agents, Sydney, Australia, November 2018 (IVA'18)*, 8 pages. https://doi.org/10.1145/3267851.3267909

IVA'18, November 5-8, 2018, Sydney, NSW, Australia

ACM 978-1-4503-6013-5/18/11$15.00
https://doi.org/10.1145/3267851.3267909

1 INTRODUCTION

Virtual agents that play the role of counselors, coaches or educators in healthcare and other fields require user retention and engagement over long periods of time in order to be effective. For example, an automated exercise coach may require a series of conversations with a user spanning months or years in duration, and intelligent tutoring systems may ultimately be designed to lead students through semester-long classes or even become life-long learning companions. Designing such systems requires approaches to maintaining user engagement over dozens, if not thousands, of interactions. Engagement is crucial, because it is typically a prerequisite for other system objectives: if a user stops interacting with a system, then it cannot have any further impact.

Previous research has explored strategies that a virtual agent application can use to boost user engagement, including variability in appearance and behavior [5], agent backstories [4], rap performance [25], co-constructed storytelling with the user [13], automatic camera motion [29], and social and relational dialogue with the agent [3]. However, these studies did not investigate the assessment and prediction of engagement in order to dynamically intervene with engagement techniques when needed to prevent users from disengaging or discontinuing use.

Several recent studies have developed and evaluated machine learning models for predicting user engagement across different non-agent applications [8,30,31]. For example, Sano et al. [31] developed a user dropout prediction model based on usage data collected from a commercial chatbot during a 2-month observation period, and achieved an accuracy of 77.6%. These existing studies, however, did not take into account behavior-related factors, which might be key in predicting long-term user engagement in behavior change interventions.

In this work, we report on the development of machine learning models that can predict future engagement for individual users in longitudinal virtual-agent based interventions, based on features derived from prior interactions. Our models were trained on 3 datasets containing usage data collected from 3 virtual agent-based health interventions, used by different user populations and deployed on different platforms (kiosk, web, mobile). In addition to basic features related to user-agent session time and frequency, we also included features related to user health behavior and dialogue

content. Evaluation results showed that our dropout prediction models achieved accuracies of 74% to 90% across the three virtual agent systems. We conclude by summarizing common factors that robustly predict user engagement or disengagement across virtual agent applications.

2 RELATED WORK

In this section, we first review existing theories related to long-term user engagement and strategies for promoting engagement with conversational agents in longitudinal behavior change interventions. We then discuss previous work on prospective user engagement prediction.

2.1 Concepts and Theories of Engagement

User engagement is crucial to any human-computer interaction. Prior HCI research has described several theories and frameworks related to both *short-term cognitive engagement* (e.g. experience of flow [14,23] and enjoyment while interacting with a technology [24]), and *long-term engagement* that spans over an extended period of time. In our current work, we focus on the latter concept of user engagement, defined as the *'duration and depth of usage'* of a system over time [12]. There are several related measures of long-term user engagement with a system in longitudinal interventions, such as: the number of voluntary interactions that users choose to have with the system in a given time period, the length of time they adhere to the system recommendations, or retention (i.e. the number of users who complete an intervention) [5]. In the context of learning activities, engagement can also be measured in terms of the learner's behavioral involvement and emotional reactions [32].

Long-term and meaningful engagement with the system has been shown to strongly impact users' behavior change in Digital Behavior Change Interventions (DBCIs) [22]. Perski et al. [26] presented a conceptual framework in which long-term engagement in DBCIs is directly influenced by factors such as the delivery, context, and setting of the intervention. Bickmore and Picard [6] proposed a theory based on a personal relationship model for promoting long-term engagement. Their theory highlights four factors that can influence long-term engagement with a system positively and negatively. According to this theory, the user's perception of their benefits during the interaction, and the user's perception of their investment in the system may increase the user's commitment to the system, while the user's perceived cost, and their perception of other applicable alternatives to the system may negatively influence their engagement in the long-term.

2.2 Maintaining Long-term User Engagement

Researchers in HCI and behavioral science have described several factors, such as goal setting [34], reminders [20], feedback [27] and rewards [17], that are associated with higher engagement in longitudinal interventions. In the field of conversational agents, social dialogue [11] and games [15] have been shown to positively impact user engagement.

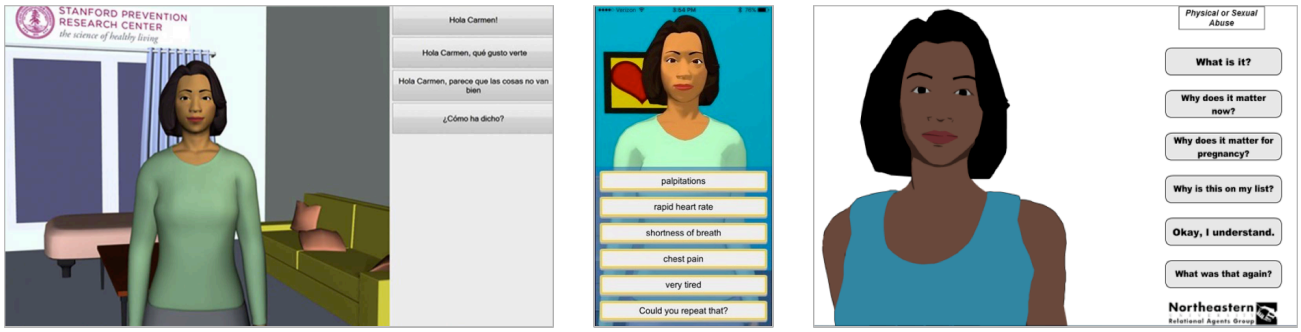
Relational agents [1] have also been shown to be an effective medium for improving long-term engagement in systems designed to promote behavior changes. Relational agents are conversational virtual characters designed to form long-term social-emotional relationships with their users [6]. These relational agents have been used as virtual coaches to motivate and guide users through a behavior change process in different contexts [18,21,28]. In a longitudinal study of physical activity promotion, Bickmore et al. [2] showed that increasing the agent's dialogue and appearance variability led to higher user engagement. Another study also demonstrated the positive effects of giving a human back story to the agent on user engagement [4]. In addition, Battaglino and Bickmore [13] conducted a study to examine the effects of co-constructed social storytelling on user engagement with conversational agents. Results of the study indicated that engagement could be improved by allowing users to contribute to a story through meaningful questions. Another approach to promoting user engagement is to treat it as a behavioral variable. For example, Bickmore et al. [7] showed that patients interacted with a system more often when the agent reminded them about the importance of frequent interactions and provided personalized feedback based on their interaction frequency.

2.3 Prospective Engagement Prediction

Recent studies have developed and evaluated machine learning models for predicting user's continued engagement with technology over time in different domains, from intelligent virtual assistants [31] to online MOOCs [8] and only health forums [30]. Of most relevance to our present study is Sano et al.'s work on predicting prospective user engagement with virtual assistants [31]. Using a large dataset of 4-month user logs of 348,295 users with a commercial intelligent assistant, the researchers developed two models to predict how frequently a user will engage with the assistant in the future given their past dialogue history. The models were trained using 338 features related to user utterance frequency, system response frequency, interaction time interval, and user profile. Evaluation results showed that their dropout prediction model could predict whether users would stop using the assistant with 77.6% accuracy after observing 2 months of their dialogue activities. In our work, we adopted several features used in Sano et al.'s models, while adding new dialogue and behavioral features specific to virtual agent-based health behavior change interventions.

Although it did not feature a virtual agent, Sadeque et al. [30] conducted a related study on predicting continued participation in online health forums. Using a large dataset collected from a support group-based social networking site, the researchers developed models that observe user activities for one month and could predict whether a user will continue participating in the support group in the future with 83% accuracy.

While not directly targeting user engagement. Kiseleva et al. [19] and other researches [16] developed models to predict *user satisfaction* with intelligent assistants, a factor that is a precursor to user engagement. Using a variety of interaction signals,



a. Carmen b. Tanya c. Gabby

Figure 1: Three agent systems used for engagement prediction: (a) desktop-based agent for physical activity promotion; (b) smartphone-based agent for atrial fibrillation; (c) web-based agent for preconception care.

including voice commands, physical touch gestures and reading patterns, Kiseleva et al. [19] showed that their model achieved 81% prediction accuracy of user satisfaction. These studies, however, differ from our work in their focus on open-ended, dialogue-based search tasks.

3 ENGAGEMENT PREDICTION MODELS

In this section, we first introduce three conversational agent systems from which we obtained training data for engagement prediction, before describing our prediction models.

3.1 Conversational Agents for Healthcare

Our engagement prediction models were developed with user logs collected from three embodied conversational agent systems designed to deliver longitudinal interventions for different health problems. Together, these systems represent a diversity of deployment platforms, usage settings, target use frequencies, and target user populations, enabling us to develop more generalizable models.

Each system features an animated female character that communicates with the user using synthetic speech. The agent's nonverbal behavior is generated using BEAT [9], and includes facial expressions, eyebrow movement, head nods, directional gazes, as well as a range of iconic, emblematic and deictic gestures. Human-agent dialogues are scripted using a custom hierarchical transition network-based scripting language. User input to the conversation is obtained via multiple choice selection of utterance options, updated at each turn of dialogue.

3.1.1 Carmen: Agent for Physical Activity Promotion

Carmen [18] is a desktop-based bilingual virtual advisor designed to promote physical activity among older Latino adults (Fig. 1a). Deployed on kiosks at five community centers in California, Carmen engages users in a 12-month intervention during which she provides individually tailored counseling on physical activity. During the intervention, users are instructed to wear a pedometer that tracks their daily walking steps. A typical session with Carmen includes: (1) greetings; (2) social chat; (3) review of walking steps since the last session based on the pedometer data; (4) personalized feedback on the user's progress in reference to their current walking goal; (5) goal setting for the period between the current and the next session; and (6) educational content. Carmen can talk with the user in either English or Spanish. The intended session frequency for this program is once-per-week for the first 2 months and twice-per-month for the remaining 10 months [18].

We obtained user logs of 114 older adults who were enrolled in a 12-month randomized controlled trial to evaluate the efficacy of the intervention. We extracted usage data of the first 8 weeks of interaction for each participant and used this dataset to develop our engagement prediction models. This first 8-week period was a crucial period of the intervention in which participants were expected to interact frequently with Carmen. On average, participants had 5.82 (SD=3.03) sessions with Carmen within the first 8 weeks, falling behind the intended weekly session schedule.

3.1.2 Gabby: Agent for Preconception Care

Gabby [28] is a web-based virtual agent designed to provide preconception care to young African American women in a 12-month intervention (Fig. 1c). Prior to interacting with Gabby, users complete a survey questionnaire to identify their preconception care risks from a list of 108 possible risk factors, ranging from substance abuse and domestic violence to nutrition and exercise. During each session with Gabby, users are guided to select the identified risks that they want to discuss. For each risk, the agent explains the importance of the risk to pregnancy and recommends actions to address it. The intervention is designed for a recommended weekly session over the course of one year [28].

We collected user logs of 201 users who participated in a 12-month randomized controlled trial to evaluate Gabby. As with Carmen, we extracted usage data of the first 8 weeks for each participant to develop our engagement prediction models. In this period, participants had an average of 2.89 sessions (SD=2.26) with Gabby, well below the recommended frequency of one session per week.

3.1.3 Tanya: Agent for Atrial Fibrillation

Tanya [21] is a smartphone-based agent that provides counseling to patients with atrial fibrillation (AF) in a 30-day intervention (Fig. 1b). The agent is designed to be used in conjunction with a mobile heart rhythm monitor that enables users to determine if they are in AF or not. During the conversation, the agent covers topics related to AF education, symptom education and reporting, medication adherence, proactive self-care, and quality of life assessment. She also provides instructions on how to use the heart rhythm monitor and promotes adherence to daily heart rhythm monitoring.

We obtained user logs of 61 users who evaluated Tanya in a 30-day randomized controlled study, and used data of the first 2 weeks of interaction for each participant for our engagement prediction models. Compared to Carmen and Tanya, participants interacted with Tanya much more frequently in shorter sessions, averaging 11.25 (SD=5.72) sessions within the first two weeks.

3.2 Engagement Prediction Tasks

Given a user's history of past interactions with an agent within an initial *observation period* (e.g. the first 4 weeks), our task is to predict whether the user will continue being engaged with the agent in the *prediction period* (e.g. the next 4 weeks). Given the 8-week user logs of Carmen and Gabby, we used the interaction history of the *first 4 weeks* to predict a user's engagement level in the *next 4 weeks*. The length of our observation period is similar to that used in previous work on prospective user engagement [30]. For Tanya, as the intervention only lasted in 30 days, we used data of the *first week* to predict a user's engagement level in the *second week*. Similar to Sano et al.'s work [31], we further broke down our task into two sub-tasks and developed separate classification models for each sub-task.

Table 1: Summary of the interventions and prediction tasks

	Carmen	Gabby	Tanya
Behavior	Physical activity	Preconception care	AF management
Language	English Spanish	English	English
Intervention duration	12 months	12 months	30 days
Recommended use frequency	Weekly	Weekly	Daily
Observation period	First 4 weeks	First 4 weeks	First week
Prediction period	Second 4 weeks	Second 4 weeks	Second week
Moderately engaged	1-3 sessions	1-3 sessions	1-6 sessions
Highly engaged	>=4 sessions	>=4 sessions	>=7 sessions

3.2.1 Dropout Prediction

This is a binary classification task in which we aim to predict if a user will stop interacting with the agent completely during the prediction period. Predicting user dropout is a crucial task, because if the user stops interacting with the agent, the intervention cannot have any further impact. Figure 2 shows the dropout rates of the three agent systems. The smartphone agent, Tanya, had the smallest dropout rate of 13.1% in the second week of the intervention. The dropout rate for Carmen was 19.3%, while Gabby had the highest dropout rate of 56.2% after the first 4 weeks of interaction.

3.2.2 Engagement Level Prediction

This is a 3-class classification task in which we further classify users into three categories: *dropout, moderately engaged* and *highly engaged* users. A user is classified as 'moderately engaged' if he/she still continues working with the agent, but does not interact with the agent as frequently as recommended during the prediction period. In contrast, a user is classified as 'highly engaged' if he/she interacts with the agent at least as frequently as recommended. Having a finer-grained classification of user engagement enables intervention designers to further personalize their engagement boosting strategies. For example, if a system is designed to proactively send reminders to disengaged users, these reminders can be sent at different frequencies depending on whether the user is likely to stop using the system completely or is just moderately disengaged.

With Carmen and Gabby, the recommended engagement in the first 8 weeks was one weekly session. Thus, we considered users as 'moderately engaged' if they had between 1-3 sessions in the second 4 weeks of the intervention (i.e. the prediction period), while 'highly engaged' users had at least 4 sessions with the agent within the prediction period. With Tanya, we classified users as 'moderately engaged' if they had between 1-6 sessions (i.e. less than one daily session), while 'highly engaged' users had at least 7 sessions within the second week of the intervention (i.e. the prediction period in Tanya).

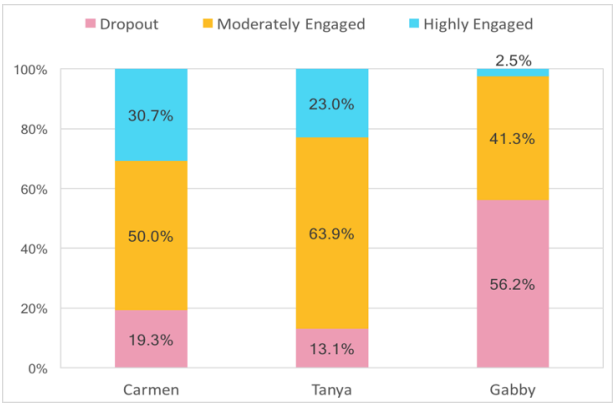


Figure 2: Distribution of users across three engagement levels for the three agent systems.

Figure 2 shows the number of 'moderately engaged' and 'highly engaged users' in the three agent systems. The percentages of 'highly engaged' users in these systems were: 2.5% for Gabby, 30.7% for Carmen, and 25% for Tanya. Table 1 provides a summary of the three agent-based health interventions and our prediction tasks.

3.3 Features

We computed a total of 13 features in three categories during the observation period: session frequency, behavior, and dialogue features. The session frequency features were available in all the three agent systems. The behavior and dialogue features were only available in Carmen.

3.3.1 Session Frequency

This category consists of 5 features capturing the frequency and duration of sessions that the user had with an agent:

Sessions: the total number of sessions with the agent in the observation period.

Average Session Duration: the average duration of sessions (in minutes) in the observation period.

First Session Duration: the duration of the first session (in minutes) in the intervention. We hypothesize that the first encounter with the agent is especially important, because it is when users form their first impressions of the agent and have an initial idea of the time and investment required in the intervention, as suggested in [33].

Min Days: minimum number of days between two consecutive sessions in the observation period.

Max Days: maximum number of days between two consecutive sessions in the observation period.

Average days: average number of days between two consecutive sessions in the observation period.

3.3.2 Behavior

This category contains 2 features that are specific to health behavior change interventions. Behavioral involvement has been considered a key component of engagement [32]. In this work, we hypothesize that there is a positive correlation between a user's positive change in their health behavior and their engagement with the agent. We were only able to compute these features from Carmen's user logs, and thus these features were specific to Carmen:

Average Steps: the average number of daily walking steps in the observation period, as recorded by the pedometer.

Percent Goal Met: the percentage of days that the user met their daily walking goal in the observation period.

3.3.3 Dialogue

This category contains 5 features related to the dialogue content and user input in their conversations with the agent. As with the behavior features, we were only able to compute these features for Carmen:

User Utterances: the total number of user utterances in the observation period.

Average Utterance Length: the average length (in words) of user utterances in the observation period.

Percent User Questions: the percentage of user utterances that are formulated as a question to the agent (e.g. 'Where are you from, Carmen?'). We hypothesize that higher percentages of user questions reflects higher levels of co-construction in the human-agent dialogue, which may lead to increased user engagement, as suggested in [13].

Percent Repeats: the percentage of user utterances that are requests for the agent to repeat her speech (i.e. when the user says the 'Could you repeat that please?' or similar options). A high percentage of repeat requests may indicate difficulty in understanding the agent's speech, which may lead to decreased user engagement.

Social Chat: the total number of social chat topics that the user has with the agent in the observation period. Social chats serve to build interpersonal relationship between the user and the agent, and thus can lead to increased long-term user engagement, as suggested in [11] [5].

For all of these features, we applied log transformation for highly skewed features, and scaled them according to the interquartile range to ensure all the features are on comparable scales.

3.4 Model Training and Evaluation

For each of the three agent systems, we developed a binary classification model and a 3-class classification model for the two engagement prediction tasks.

We first divided each dataset into training and test sets with the ratio of 7:3. To address the problem of imbalanced classes in our datasets, we oversampled the minority classes in our training set using the synthetic minority over-sampling technique (SMOTE) [10]. We experimented with three classification algorithms: SVM, Random Forest, and Gradient Boosting classifiers. For each classifier, we tuned a set of hyperparameters on our training set using Grid Search with 5-fold cross-validation. Gradient boosting generally performed the best among the three methods with our datasets, and thus we used it to train all our final models. We tuned 5 hyperparameters of our gradient boosting classifiers, including: number of tree estimators, learning rate, maximum depth of individual tree estimators, minimum number of samples required to split an internal node in a tree, and minimum number of samples required to be a leaf node in a tree.

For Carmen, in addition to models trained on all 13 features, we also trained separate models for each of the three feature categories (session frequency, behavior, and dialogue). This allows us to examine the effectiveness of each feature category in predicting prospective user engagement.

We evaluated the performance of our classification models on our held-out test set, using four metrics: accuracy, macro-averaged F1, macro-averaged precision, and macro-averaged recall. For baseline comparison, we used a naïve classifier which always predicts the majority class.

Table 2: Performance of the dropout and engagement level prediction models trained using the Session Frequency features only

		Dropout Prediction				Engagement Level Prediction			
		Accuracy	**F1**	**Precision**	**Recall**	**Accuracy**	**F1**	**Precision**	**Recall**
Carmen	Baseline	0.807	0.447	0.404	0.5	0.5	0.222	0.167	0.333
	Session Frequency	0.743	0.65	0.657	0.645	0.543	0.534	0.555	0.564
Gabby	Baseline	0.639	0.390	0.320	0.5	0.562	0.428	0.461	0.530
	Session Frequency	0.738	0.729	0.728	0.745	0.656	0.454	0.493	0.560
Tanya	Baseline	0.869	0.465	0.434	0.5	0.639	0.26	0.213	0.333
	Session Frequency	0.895	0.842	0.842	0.842	0.629	0.631	0.651	0.658

4 RESULTS AND DISCUSSION

In this section, we first report results of our engagement prediction models trained using the *Session Frequency* feature category only, which was available in all the three agent systems. We then present results of our models trained using the entire feature set on the Carmen dataset.

4.1 Session Frequency-based Models

Table 2 shows the performance results of our dropout and engagement level classification models trained using the Session Frequency sessions only. All models outperformed the naïve classifier in terms of precision and recall. The accuracy of the dropout prediction models ranged from 0.743 (F1=0.65) in Carmen to 0.738 (F1=0.729) in Gabby and 0.895 (F1=0.842) in Tanya. The accuracy of our 3-class engagement level prediction models ranged from 0.543 (F1=0.534) in Carmen to 0.656 (F1=0.454) in Gabby and 0.629 (F1=0.631) in Tanya. In all our models, the *Sessions* feature (i.e. the total number of sessions that user had with an agent during the observation period) was shown to be the most important feature.

4.2 Models Using the Entire Feature Set

4.2.1 Model Performance

Table 3 presents the performance results of our dropout and engagement level prediction models trained on the Carmen dataset using each of the three feature categories and a combination of all features. All models outperformed the baseline classifier in terms of precision and recall. Out of the three categories, behavior features were shown to be the most effective in predicting prospective engagement. For dropout prediction, the model trained only with behavior features achieved an accuracy of 0.8 (F1=0.728). Our final model trained on the entire feature set performed the best, with an accuracy of 0.829 (F1=0.757) for dropout prediction and 0.66 (F1=0.66) for engagement level prediction. For comparison, Sano et al.'s dropout prediction model (which was trained on a much bigger dataset using a set of 338 features and a longer observation period of 8 weeks) achieved an accuracy of 0.776 [31].

4.2.2 Feature Importance

Figure 3 presents the most important features with large weights (>=0.1) learned by our dropout and engagement level prediction models when trained on the Carmen dataset with the entire feature set. The top 6 most important features in the dropout prediction models included 4 dialogue features (number of user utterances, percentage of user questions, percentage of repeat requests, and number of social chat topics completed), along with the number of sessions and the percentage of daily goals met by the user in the observation period. The top 4 most important features with weight >=0.1 in our engagement level prediction model included the number of sessions, the number of user utterances, the percentage of daily goals met and the percentage of repeat requests. Results of Pearson's correlation analysis showed that there were significant, positive correlations

Table 3: Performance of the dropout and engagement level prediction models trained on Carmen using different feature categories

	Dropout Prediction				Engagement Level Prediction			
	Accuracy	**F1**	**Precision**	**Recall**	**Accuracy**	**F1**	**Precision**	**Recall**
Baseline	0.807	0.447	0.404	0.5	0.5	0.222	0.167	0.333
Session Frequency	0.743	0.65	0.657	0.645	0.543	0.534	0.555	0.564
Behavior	0.8	0.728	0.7384	0.72	0.571	0.58	0.614	0.601
Dialogue	0.743	0.65	0.657	0.645	0.543	0.542	0.552	0.601
All	**0.829**	**0.757**	**0.786**	**0.739**	**0.66**	**0.66**	**0.683**	**0.678**

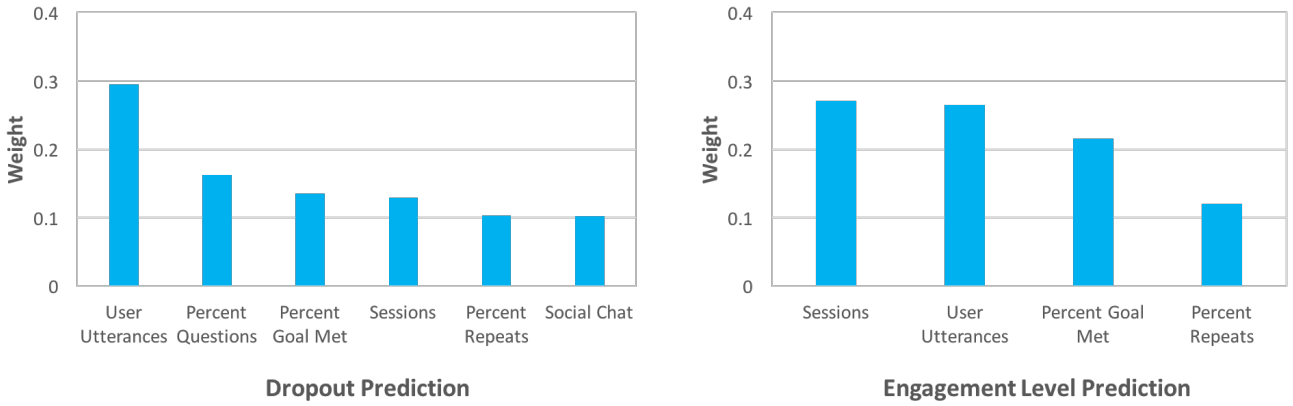


Figure 3: Most important features with large weights in the dropout and engagement level prediction models.

between the number of sessions in the prediction period and *number of sessions*, the *number of user utterances*, the *number of social chats*, the *percentage of user questions*, and *the percentage of daily goals met* in the observation period.

4.2.3 Effects of Observation Length

To examine the effect of the observation length on prediction performance, we varied the observation period from 1 week to 4 weeks while maintaining the prediction period as the next 4 weeks following the observation period.

Figure 4 presents the accuracy and F1 score of our dropout and engagement level prediction models across 4 different observation lengths for the Carmen dataset. Increasing the observation length from 1 week to 4 weeks led to 56% improvement in F1 score for dropout prediction and 73.8% improvement in F1 score for engagement level prediction. A substantial increase in the prediction performance was achieved when moving from 2-week to 3-week observation, especially for the dropout prediction task. Using data from the 3-week observation, the dropout prediction model achieved a reasonably good performance (accuracy=0.857, F1=0.730). This suggests that we could potentially start detecting user dropout after the first three weeks of an intervention instead of waiting for the full 4 weeks, for interventions lasting 5 weeks or longer.

5 CONCLUSIONS

We reported the results of machine learning–based models that can predict user engagement in longitudinal virtual agent-based health interventions. Prediction accuracies are very good for predicting dropouts, and moderately good for predicting more fine-grained classes of user engagement.

Our finding that several dialogue features are important in predicting engagement indicates that several aspects of health counseling dialogue are not only important for user satisfaction and health outcomes, but for maintaining long-term retention as well. The importance of social dialogue, in particular, reinforces earlier findings that social, "off task" talk plays an important role in many task-oriented conversations [1,2]. While our datasets were all from virtual agent-based health interventions, we believe our methods and findings generalize to a much wider class of longitudinal interventions.

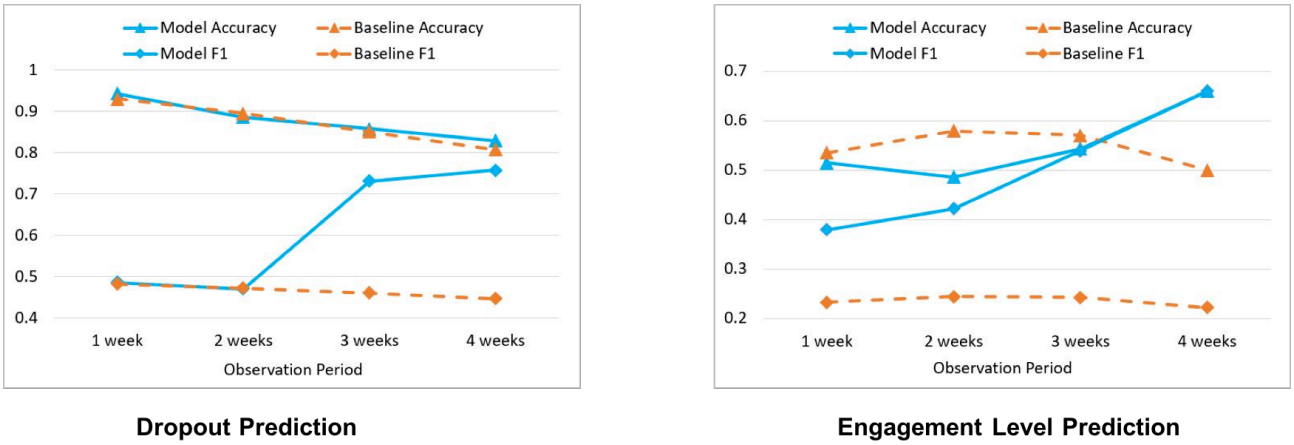


Figure 4: Performance of the dropout and engagement level prediction models across different observation lengths. The performance of the baseline models is included for comparison.

Future work includes the development of more accurate models, and the design of re-engagement interventions a virtual agent can use when user dropout is predicted, taking into account all that is known about the user's personal characteristics and discourse and interaction history.

ACKNOWLEDGMENTS

This work is supported in part by the US National Institutes of Health under grant R01HL116448 and the Doris Duke Charitable Foundation under grant 2015084.

REFERENCES

[1] Timothy W. Bickmore and Justine Cassell. 2001. Relational Agents: A Model and Implementation of Building User Trust. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '01), 396–403. DOI:https://doi.org/10.1145/365024.365304

[2] Timothy W. Bickmore, Amanda Gruber, and Rosalind Picard. 2005. Establishing the computer-patient working alliance in automated health behavior change interventions. *Patient Educ Couns* 59, 1, 21–30. DOI:https://doi.org/10.1016/j.pec.2004.09.008

[3] Timothy W. Bickmore, Laura Pfeifer, and Daniel Schulman. 2011. Relational Agents Improve Engagement and Learning in Science Museum Visitors. In *Proceedings of the 11th International Conference on Intelligent Virtual Agents*, 55–67. DOI:https://doi.org/10.1007/978-3-642-23974-8_7

[4] Timothy W. Bickmore, Daniel Schulman, and Langxuan Yin. 2009. Engagement vs. Deceit: Virtual Humans with Human Autobiographies. In *Proceedings of the 9th International Conference on Intelligent Virtual Agents* (IVA '09), 6–19. DOI:https://doi.org/10.1007/978-3-642-04380-2_4

[5] Timothy Bickmore, Daniel Schulman, and Langxuan Yin. 2010. Maintaining Engagement in Long-term Interventions with Relational Agents. *Appl Artif Intell* 24, 6 (July 2010), 648–666. DOI:https://doi.org/10.1080/08839514.2010.492259

[6] Timothy W. Bickmore and Rosalind W. Picard. 2005. Establishing and Maintaining Long-term Human-computer Relationships. *ACM Trans. Comput.-Hum. Interact.* 12, 2 (June 2005), 293–327. DOI:https://doi.org/10.1145/1067860.1067867

[7] Timothy W. Bickmore, Kathryn Puskar, Elizabeth A. Schlenk, Laura M. Pfeifer, and Susan M. Sereika. 2010. Maintaining reality: Relational agents for antipsychotic medication adherence. *Interacting with Computers* 22, 4 (July 2010), 276–288. DOI:https://doi.org/10.1016/j.intcom.2010.02.001

[8] Carolyn P. Rosé and George Siemens. 2014. Shared Task on Prediction of Dropout Over Time in Massively Open Online Courses - Semantic Scholar. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 39–41.

[9] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2001. BEAT: the Behavior Expression Animation Toolkit. 477–486. In *Proceedings of SIGGRAPH 2001*. DOI:https://doi.org/10.1145/383259.383315

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *1* 16, 321–357. DOI:https://doi.org/10.1613/jair.953

[11] Miguel Coronado, Carlos A. Iglesias, Álvaro Carrera, and Alberto Mardomingo. 2018. A cognitive assistant for learning java featuring social dialogue. *International Journal of Human-Computer Studies*. DOI:https://doi.org/10.1016/j.ijhcs.2018.02.004

[12] Mick P. Couper, Gwen L. Alexander, Nanhua Zhang, Roderick J. A. Little, Noel Maddy, Michael A. Nowak, Jennifer B. McClure, Josephine J. Calvi, Sharon J. Rolnick, Melanie A. Stopponi, and Christine Cole Johnson. 2010. Engagement and retention: measuring breadth and depth of participant use of an online intervention. *J. Med. Internet Res.* 12, 4 (November 2010), e52. DOI:https://doi.org/10.2196/jmir.1430

[13] Cristina Battaglino and Timothy Bickmore. 2015. Increasing the Engagement of Conversational Agents through Co-Constructed Storytelling. In *Proceedings of the Eighth Workshop on Intelligent Narrative Technologies*.

[14] Juho Hamari, David J. Shernoff, Elizabeth Rowe, Brianno Coller, Jodi Asbell-Clarke, and Teon Edwards. 2016. Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning. *Computers in Human Behavior* 54, 170–179. DOI:https://doi.org/10.1016/j.chb.2015.07.045

[15] Hayato Kobayashi, Kaori Tanio, and Manabu Sassano. 2015. Effects of Game on User Engagement with Spoken Dialogue System. In *Proceedings of the SIGDIAL*.

[16] Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. 2015. Automatic Online Evaluation of Intelligent Assistants. In *Proceedings of the 24th International Conference on World Wide Web* (WWW '15), 506–516. DOI:https://doi.org/10.1145/2736277.2741669

[17] Zarnie Khadjesari, Elizabeth Murray, Eleftheria Kalaitzaki, Ian R. White, Jim McCambridge, Simon G. Thompson, Paul Wallace, and Christine Godfrey. 2011. Impact and costs of incentives to reduce attrition in online trials: two randomized controlled trials. *J. Med. Internet Res.* 13, 1, e26. DOI:https://doi.org/10.2196/jmir.1523

[18] Abby C. King, Ines Campero, Jylana L. Sheats, Cynthia M. Castro Sweet, Dulce Garcia, Aldo Chazaro, German Blanco, Michelle Hauser, Fernando Fierros, David K. Ahn, Jose Diaz, Monica Done, Juan Fernandez, and Timothy Bickmore. 2017. Testing the comparative effects of physical activity advice by humans vs. computers in underserved populations: The COMPASS trial design, methods, and baseline characteristics. *Contemp Clin Trials* 61, 115–125. DOI:https://doi.org/10.1016/j.cct.2017.07.020

[19] Julia Kiseleva, Kyle Williams, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Predicting User Satisfaction with Intelligent Assistants. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '16), 45–54. DOI:https://doi.org/10.1145/2911451.2911521

[20] Haotian Lin and Xiaohang Wu. 2014. Intervention strategies for improving patient adherence to follow-up in the era of mobile information technology: a systematic review and meta-analysis. *PLoS ONE* 9, 8, e104266. DOI:https://doi.org/10.1371/journal.pone.0104266

[21] Jared W. Magnani, Courtney L. Schlusser, Everlyne Kimani, Bruce L. Rollman, Michael K. Paasche-Orlow, and Timothy W. Bickmore. 2017. The Atrial Fibrillation Health Literacy Information Technology System: Pilot Assessment. *JMIR Cardio* 1, 2, e7. DOI:https://doi.org/10.2196/cardio.8543

[22] Susan Michie, Lucy Yardley, Robert West, Kevin Patrick, and Felix Greaves. 2017. Developing and Evaluating Digital Interventions to Promote Behavior Change in Health and Health Care: Recommendations Resulting From an International Workshop. *J. Med. Internet Res.* 19, 6, e232. DOI:https://doi.org/10.2196/jmir.7126

[23] Jeanne Nakamura and Mihaly Csikszentmihalyi. 2014. The Concept of Flow. In *Flow and the Foundations of Positive Psychology*. Springer, Dordrecht, 239–263. DOI:https://doi.org/10.1007/978-94-017-9088-8_16

[24] Heather L. O'Brien and Elaine G. Toms. What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology* 59, 6, 938–955. DOI:https://doi.org/10.1002/asi.20801

[25] Stefan Olafsson, Everlyne Kimani, Reza Asadi, and Timothy Bickmore. 2017. That's a Rap: Increasing Engagement with Rap Music Performance by Virtual Agents. In *Proceedings of 17th International Conference on Intelligent Virtual Agents*, 325–334. DOI:https://doi.org/10.1007/978-3-319-67401-8_41

[26] Olga Perski, Ann Blandford, Robert West, and Susan Michie. 2017. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. *Behav. Med. Pract. Policy Res.* 7, 2 (June 2017), 254–267. DOI:https://doi.org/10.1007/s13142-016-0453-1

[27] Christopher Peters, Ginevra Castellano, and Sara de Freitas. 2009. An Exploration of User Engagement in HCI. In *Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots*, 9:1–9:3.

[28] Jingjing Ren, Timothy Bickmore, Megan Hempstead, and Brian Jack. 2014. Birth Control, Drug Abuse, or Domestic Violence? What Health Risk Topics Are Women Willing to Discuss with a Virtual Agent? In *Proceedings of the 14th International Conference on Intelligent Virtual Agents*, 350–359. DOI:https://doi.org/10.1007/978-3-319-09767-1_46

[29] Lazlo Ring, Dina Utami, Stefan Olafsson, and Timothy Bickmore. 2016. Increasing Engagement with Virtual Agents Using Automatic Camera Motion. In *Proceedings of the 16th International Conference on Intelligent Virtual Agents*, 29–39. DOI:https://doi.org/10.1007/978-3-319-47665-0_3

[30] Farig Sadeque, Thamar Solorio, Ted Pedersen, Prasha Shrestha, and Steven Bethard. 2015. Predicting Continued Participation in Online Health Forums. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, 12–20.

[31] Shumpei Sano, Nobuhiro Kaji, and Manabu Sassano. 2016. Prediction of Prospective User Engagement with Intelligent Assistants. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 1203–1212.

[32] Ellen A Skinner and Michael J Belmont. 1993. Motivation in the Classroom: Reciprocal Effects of Teacher Behavior and Student Engagement Across the School Year. *Journal of Educational Psychology* 85, (1993), 11.

[33] Laura Vardoulakis. 2013. Social Desirability Bias and Engagement in Systems Designed for Long-Term Health Tracking. *PhD Dissertation*.

[34] Anna Weston, Leanne Morrison, Lucy Yardley, Max Van Kleek, and Mark Weal. 2015. Measurements of engagement in mobile behavioural interventions? In *Proceedings of Digital Health 2015*.