



Predicting and Exploring Abandonment Signals in a Banking Task-Oriented Chatbot Service

Chieh Hsu, Hsin-Chien Tung, Hong-Han Shuai & Yung-Ju Chang

To cite this article: Chieh Hsu, Hsin-Chien Tung, Hong-Han Shuai & Yung-Ju Chang (20 Nov 2023): Predicting and Exploring Abandonment Signals in a Banking Task-Oriented Chatbot Service, International Journal of Human-Computer Interaction, DOI: [10.1080/10447318.2023.2282220](https://doi.org/10.1080/10447318.2023.2282220)

To link to this article: <https://doi.org/10.1080/10447318.2023.2282220>



Published online: 20 Nov 2023.



Submit your article to this journal [↗](#)



Article views: 206



View related articles [↗](#)





View Crossmark data [↗](#)



Citing articles: 1 View citing articles [↗](#)

Predicting and Exploring Abandonment Signals in a Banking Task-Oriented Chatbot Service

Chieh Hsu^a , Hsin-Chien Tung^a , Hong-Han Shuai^b , and Yung-Ju Chang^a 

^aDepartment of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan; ^bDepartment of Electrical Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

ABSTRACT

In this study, we developed predictive models to address the problem of chatbot abandonment, a problem that can result in losing business opportunities. Specifically, we target on a conversation log dataset of a banking chatbot involving 1,373 users and hand-crafted features. By leveraging a pre-trained BERT model on the textual features and the hand-crafted features, the model achieved an F1-score of 0.89 in predicting discontinued conversation and 0.80 in predicting abandonment. Our findings indicate that textual features help capture more abandonment, while hand-crafted features improve detection precision. Our analysis with SHAP and LIME revealed that user typing, the chatbot expressing inability of recognizing intent, and the chatbot asking what users want to do during an ongoing conversation are top signals of user abandoning the chatbot. These findings suggest that chatbot designers should consider providing pre-set options or constraints for user inputs and presenting possible intents to the user and avoid expressing inability, incompetence, or ignoring the users' current attempt.

KEYWORDS

Task-oriented chatbot; machine learning; BERT; explainable AI; abandonment; conversation breakdown

1. Introduction

Conversational User Interfaces (CUI) have revolutionized the way humans interact with computers by providing a simpler and more natural way to communicate through natural language. One of the most significant implementations of CUIs is text-based chatbots, which have been increasingly applied in various domains, including education (Farah et al., 2022; Kane, 2016; Smutny & Schreiberova, 2020), healthcare (Amer et al., 2021; Viswanath Prakash & Das, 2020), answering queries (Athreya et al., 2018) and presenting specific information (Kannagi et al., 2018; Segura et al., 2019; Vassos et al., 2016) etc.

In recent years, utilization of task-oriented chatbots in businesses has increased due to their cost-effective and efficient means of providing customer services (Atif et al., 2021; Buhalis & Cheng, 2020; Handoyo et al., 2018; Prasetyo et al., 2020). This has allowed many online services to provide 24/7 service to a large number of customers without high human resource costs, making them an attractive alternative or primary means of service provision. In particular, financial institutions have increasingly utilized task-oriented chatbots as a way to complement their existing customer service channels (Atif et al., 2021; Khan & Rabbani, 2021; Okuda & Shoda, 2018; Tenemaza et al., 2020). According to survey and analysis (Bhaskaran, 2020), the banking industry could potentially save up to 7.3 billion dollars in costs and 862 million working hours by 2023 through chatbot adoption. As the adoption of chatbots in financial institutions becomes

more widespread, there are numerous researches exploring human's perceptions and satisfaction with banking chatbots (Bagana et al., 2021; Bouhia et al., 2022; Eren, 2021; Mulyono & Sfenrianto, 2022). Eren (2021) involved 240 customers who used chatbots for banking transactions and found that perceived performance, perceived trust, and corporate reputation significantly influence customer satisfaction.

However, conversation breakdowns can lead to user frustration (Ashktorab et al., 2019; Engelhardt et al., 2017) and abandonment of the chatbot service (Li et al., 2020), hindering the potential benefits of chatbot adoption. While taking over conversations with customer service staffs is one option to mitigate this, it risks overwhelming businesses and contradicts their goal of reducing human effort. Previous research by Li et al. (2020) presented several such signals, but lacked computational support for correlation between these signals and abandonment. To address this gap, our research questions are as follows:

RQ1: How accurately can we predict user abandonment using machine learning and hand-crafted features in conversation logs?

RQ2: What features are important for predicting user abandonment that can serve as signals of abandonment?

RQ3: Can textual features provide additional signals of user abandonment and improve prediction performance?

The study used a conversation log of 1373 users and hand-crafted features from a banking chatbot (Li et al., 2020)

to train models for predicting user abandonment. The results showed that a pre-trained BERT model achieved an F1-score of 0.89 in predicting discontinued conversation and 0.80 in predicting abandonment. Textual features were found to be helpful in capturing more abandonment, while hand-crafted features improved detection precision. Visual analysis using Explainable Artificial Intelligence (XAI) tools, SHAP and LIME, identified user typing, chatbot inability to recognize intent, and chatbot asking what users want to do during an ongoing conversation as important indicators of chatbot abandonment. Chatbot designers may benefit from increasing the occurrence of presenting pre-set options or constraints for user inputs and providing possible intents to the user, while avoiding expressions of inability, incompetence, or ignoring the user's current attempt to potentially prevent abandonment. The main contributions of this paper are summarized as follows:

1. Our study demonstrates the feasibility of using machine learning and hand-crafted features in conversation logs to predict user abandonment, with a pre-trained BERT model leveraging textual features to capture more instances of abandonment at the cost of prediction precision.
2. We identify user typing, chatbot inability to recognize intent, and the chatbot asking what users want to do during an ongoing conversation as important features indicative of user abandonment.
3. We distinguish between two types of conversation breakdowns, non-recognition and mis-recognition, with non-recognition being more critical and requiring avoidance as it is more likely to lead to abandonment. We also propose a third type, restarting conversation, which is often neglected but highly predictive of chatbot abandonment.

2. Related work

2.1. Human-Chatbot interaction

The interaction between humans and computers has received significant research attention. Researchers have been exploring ways to understand user behavior and perceptions in using Conversational User Interfaces (CUI) to improve the process of human-computer interaction, making it smooth and pleasant. Many studies related to Voice Assistants (VA) are driven by this motivation (Acikgoz & Vega, 2022; da Silva et al., 2022; Hong & Cho, 2022; Lee et al., 2022; Nguyen et al., 2019). For instance, da Silva et al. (2022) conducted experiments with literate and illiterate individuals using Google Assistant and identified two essential characteristics that need improvement in the interaction with illiterates. In addition, Nguyen et al. (2019) discussed the gender effects on the relationship between trust, mobile self-efficacy, and attitude toward Voice User Interface (VUI) use. Human-chatbot interaction is also a prominent research perspective that researchers have been focusing on (Chaves & Gerosa, 2019; Følstad & Bjerkreim-Hanssen, 2023; Grudin

& Jacques, 2019; Haugeland et al., 2022; Jain et al., 2018; Weisz et al., 2019). For example, Haugeland et al. (2022) using questionnaires and interviews with 35 participants and revealed that button interaction can enhance both the pragmatic and emotional experience related to the individual's psychological well-being compared to free text interaction. They also found that topic-oriented conversation in customer service chatbots can lead to increasing human-like qualities and user experience. In their meta-analysis of 83 papers about text-based chatbot in the last 10 years of research, Rapp et al. (2021) suggest that user satisfaction is closely tied to the alignment between users' expectations of a chatbot's capabilities and their actual perceptions of its performance. Therefore, it is crucial to design chatbots that can meet users' expectations to ensure a positive user experience.

Given the variety type of chatbots, prior research has attempted to classify chatbots into different types (Følstad et al., 2019; Hussain et al., 2019). In particular, in terms of whether the chatbot is designed to allow users to accomplish specific tasks, chatbots can further be divided into task-oriented and non-task-oriented chatbots. Task-oriented chatbots are designed to help users complete specific tasks, such as booking a hotel (Buhalis & Cheng, 2020), flight tickets (Handoyo et al., 2018). However, even though these chatbots are designed to understand and process messages related to these specific tasks, which potentially limits the scope of free text input, conversations can still break down during interactions (Li et al., 2020). As it is crucial to assist users in recognizing, diagnosing, and recovering from errors to create effective usability (Amershi et al., 2019; Langevin et al., 2021; Nielsen, 1995), prior research has explored ways to repair conversation breakdowns (Ashktorab et al., 2019; Brandtzaeg & Følstad, 2017; Engelhardt et al., 2017; Langevin et al., 2021; Silva & Canedo, 2022; Yarosh et al., 2018; Yeh et al., 2022). For example, Lee et al. (2010) tested three repair strategies including apologies, compensation, and options for the user in an online study and found that people's orientation influenced which recovery strategy worked best. Specifically, people with a relational orientation responded best to an apology, whereas those with a utilitarian orientation responded best to compensation. Ashktorab et al. (2019), on the other hand, compared eight repair strategies. They found that providing options and explanations were generally favored because they showed the chatbot's initiative and were actionable to recover from breakdowns. Yeh et al. (2022) studied the effectiveness of eight combinations of two guidance types (example-based vs. rule-based) at four timings in helping users use task-oriented chatbots. They found that each guidance type and timing has its own strengths and weaknesses and thus considering when to present specific guidance is important to the user experience of chatbot interaction.

Despite the availability of different strategies to repair conversation breakdowns, there are still issues leading to user abandonment that require human intervention in chatbot services (Jain et al., 2018). Thus, identifying when users are about to abandon a chatbot and providing timely

assistance has become a critical issue. In particular, Li et al. (2020) conducted an observational study using a conversation log of users interacting with a banking chatbot. They identified 12 types of conversational non-progress and three signs that users were about to abandon the chatbot, including three consecutive instances of non-progress, consecutive use of message reformulation or switching subjects, and using message reformulation as the final strategy. However, these signs lack computational support, which makes their predictive power for identifying instances of user abandonment unclear. Unfortunately, few studies have used machine learning to explore these signs. Therefore, our study aims to examine the feasibility of predicting user abandonment of task-oriented chatbots and explore abandonment signals using XAI tools to fill this gap in the literature.

2.2. Machine learning and explainable AI (XAI) tools

Machine learning focuses on the development of algorithms and statistical models that can learn from data, without being explicitly programmed. It allows computers to automatically improve their performance on a specific task by learning from experience and is used in a wide range of applications, including image recognition (Uchida, 2013; van Heel et al., 1996), natural language processing (NLP) (Devlin et al., 2018; Lee et al., 2017; Vaswani et al., 2017; Zhou et al., 2018), and classification (Myles et al., 2004; Rigatti, 2017; Xue et al., 2009). Prior research also adopted machine learning in user abandonment predictions, for example, Wu and Zhang (2020) selected the Random Forest algorithm for predicting users' good abandonment behavior in mobile search; Williams and Zitouni (2017) modeled the good abandonment detection problem as a sequence classification problem and used Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN) for prediction.

However, machine learning models have long been regarded as a black box, of which the result and prediction outcome is challenging to interpret and make sense of. Therefore, to let people understand the reason behind the judgements from models, numerous researchers have developed various techniques to construct explainable AI (Bach et al., 2015; Ebermann et al., 2023; Nauta et al., 2019; Ribeiro et al., 2016; Zhang et al., 2018). Ribeiro et al. (2016), for example, proposed LIME (Local Interpretable Model-Agnostic Explanations) to generate human-understandable explanations for specific predictions made by a machine learning model; Lundberg and Lee (2017) proposed SHAP (SHapley Additive exPlanations) to value attribute a prediction to each feature; and Hastie et al. (2009) proposed partial dependence plots (PDP) to measure the impact of features on the model's predictions. These XAI tools are useful for understanding the behavior of complex machine learning models by explaining how individual features contribute to their predictions. For instance, in a study by Choi et al. (2020), a Long Short-Term Memory (LSTM) was used to classify job advertisements, and LIME was used to identify the cause of the classification result. They found that the word "competencies" had the most positive influence for IT

jobs, while "rest" had the most negative influence. Furthermore, in the study by Kim et al. (2022), they proposed several approaches to enhance the detection of hate speech and utilized LIME to observe the rationale behind different BERT models' judgments on the same speech utterances. In another study by Parsa et al. (2020), eXtreme Gradient Boosting (XGBoost) was used to detect accidents using real-time data, and SHAP was used to analyze the importance of individual features. They found that traffic-related features, specifically the difference in speed before and after an accident, were the most important features. Understanding how machine learning models make decisions is crucial for building trust in these models and improving their performance. XAI tools are important in achieving this goal, as they can help identify potential issues with the data and features used for training. Similarly for the current study, by using XAI tools to facilitate visual analysis, we can gain valuable insights into the instances where users abandon a chatbot service.

3. Methodology

3.1. Experiment setups

To answer our research questions, we utilized a two-step approach that included developing a model to predict chatbot abandonment and using the SHAP XAI tool to determine the features with the greatest impact on predicting whether users would abandon the chatbot or not. This was inspired by Li et al. (2020), who used hand-crafted annotations to analyze conversation breakdowns and users' coping strategies, including abandoning the chatbot. This step was used to answer the first two research questions.

Recognizing that hand-crafted features may not capture all the information indicative of user abandonment, we incorporated the BERT language model, which is effective in processing natural language. Our approach combined both hand-crafted features and text as input to predict user abandonment and aimed to leverage BERT's ability to identify more specific and fine-grained signals beyond what was identified through hand-crafted features. To further identify signals, we used the LIME XAI tool to observe the language features that the model attended to in users' inputs and the chatbot's output.

In the following sub-sections, we will provide additional details on our methodology, including a detailed description of our data processing approach.

3.2. Dataset and data preprocessing

3.2.1. Dataset: Conversation log with a banking chatbot

This study used a dataset provided by Li et al. (2020), which is a conversation log between users and a task-oriented chatbot built on Facebook Messenger of a banking institution in Taiwan, recorded from May 1 2017 to July 31 2017, provided by Li et al. (2020). The chatbot provided several banking services such as currency exchange rates, credit card overviews, investment information, etc. The data was stored

in a spreadsheet with 19,451 conversation exchanges, each represented as a separate row. This dataset has been organized, annotated, and analyzed by Li et al. to understand how users reacted to a lack of progress in their conversations with the chatbot, including abandoning the conversation (Li et al., 2020).

3.2.2. Hand-crafted features

To create our model, we utilized the annotated dataset which contained 88 hand-crafted features characterizing various feature types of each conversation exchange. We focused on the features that were most indicative of a user's decision to abandon the chatbot and used Recursive Feature Elimination (RFE) which can be used based on different algorithms (e.g., logistic regression or SVM) to eliminate the features recursively until the desired number is obtained. Mao et al. (2022) used SVM-RFE to select a subset of features suitable for loan payment prediction. In this paper, we used Logistic Regression based Recursive Feature Elimination (LR-RFE) to score the importance of the hand-crafted features and select the top 20 hand-crafted features, which are categorized into those representing the user and the chatbot. The hand-crafted features representing the user include those indicating the intention to communicate, the way users interact with the chatbot, and the content of the user's input. The hand-crafted features representing the chatbot include those indicating the chatbot's responses that have been deemed as two types of conversation breakdowns or "non-progress" (Li et al., 2020), and those related to Dialog State Tracking (DST). DST was used to determine the objective or intention of the user at a particular turn, taking into account all the previous interactions in the dialog up to that point (Henderson et al., 2014). A detailed list of the 20 selected hand-crafted features is provided in Table 1.

The example shown in Figure 1 demonstrates the DST-related features, where the chatbot is engaged in a credit card service conversation with the user. The conversation flow is initiated when the user requests the service, and the chatbot enters the DST mode and asks questions to gather information. In the final exchange, the chatbot provides the user with credit card information. The label "Chatbot in DST" is applied to all three exchanges, while "Chatbot providing DST information" is only applied to the third exchange, as in the first two exchanges, the chatbot is

requesting information from the user. Figure 1(B) illustrates an example of the chatbot starting an irrelevant DST, while Figure 1(C) shows an example of the chatbot exiting DST. The hand-crafted features associated with DST (F17–F20) can be further divided into features related to the chatbot's response (F17) and the chatbot's status (F18–F20). Thus, we obtained five types of hand-crafted features, which include the user's intention (F1–F5), the way users interact with the chatbot (F6–F7), the content of the user's input (F8–F14), the chatbot's non-progress responses (F15–F17), and the chatbot's status (F18–F20). This feature selection approach helped to simplify the model and enhance its interpretability.

3.2.3. Prediction targets

The main aim of our research was to predict instances of user abandonment in chatbot interactions. To ensure consistency and comparability with the study by Li et al. (2020) that analyzed the same dataset, we adopted their 10-day threshold to differentiate cases of user abandonment. Their threshold was established based on the observation that 79.5% of users returned to the chatbot within 10 days of leaving after experiencing a non-progress event. We defined *abandonment* as cases where a user left the chatbot for more than 10 days after a non-progress event, as annotated in the dataset. In addition, we developed a model that predicts user leaving a conversation for more than 30 min after a non-progress event, which we referred to as *discontinued conversation*. Although this type of leave may not necessarily signify abandonment, it could imply a conversation's friction, causing the user to leave, even if only temporarily. Therefore, these discontinued conversations included two types of leaving events: *abandonment* and *temporary leaving*, where the latter refers to a user leaving a conversation for more than 30 min but returning within 10 days. We adopted the 30-min threshold to identify a break in a continuous conversation session since it has been consistently recommended or utilized in previous studies, such as Catledge and Pitkow (1995); Jones and Klinkner (2008); Li et al. (2020). Thus, we grouped continuous conversation sessions based on this threshold, classifying any conversation that did not receive a user's input within 30 min after the chatbot's response as a discontinued conversation. Any subsequent user input after this threshold was considered the start of a

Table 1. Top 20 hand-crafted features from feature selection for input data.

User			
F1	User requesting information	F8	User providing information in wrong format
F2	User providing information	F9	User mis-recognition
F3	User finding assistants	F10	User providing incoherent input
F4	User complaining	F11	User giving an unfinished message
F5	User wants to go back to the previous step in the middle of the conversation	F12	User finishing an unfinished message
F6	User typing	F13	User enhancing the previous input
F7	User using button	F14	User cross-exchange response
Chatbot			
F15	Chatbot mis-recognition	F18	Chatbot in DST
F16	Chatbot non-recognition	F19	Chatbot starting irrelevant DST
F17	Chatbot providing DST information	F20	Chatbot exiting DST

DST stands for Dialog State Tracking.

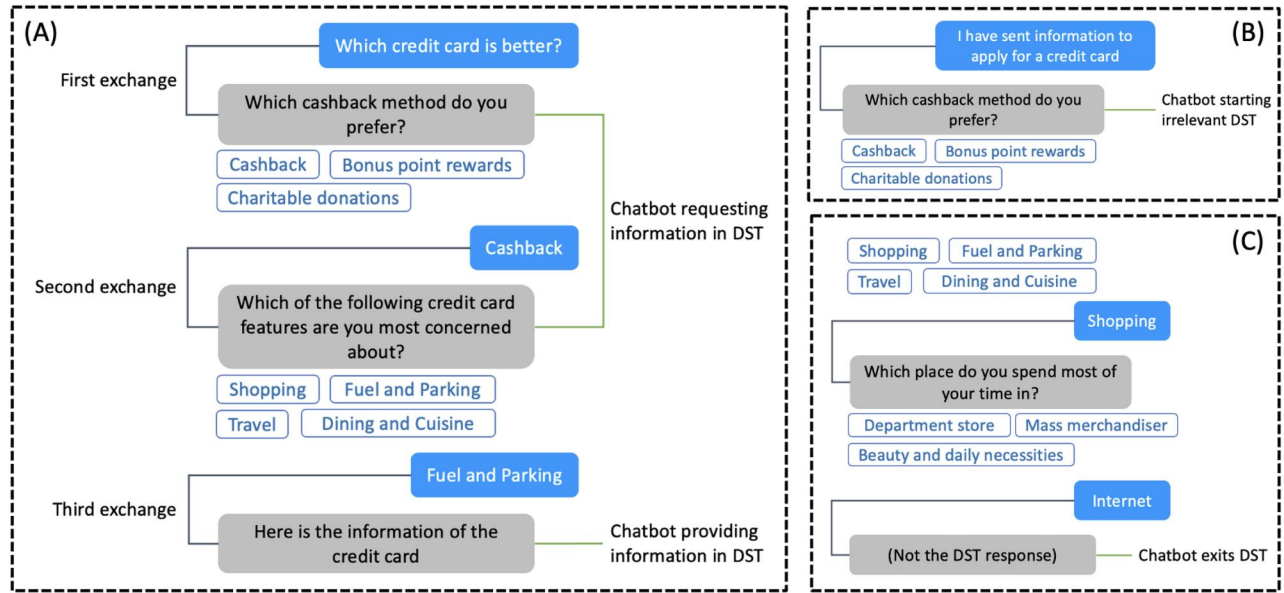


Figure 1. English translation examples of hand-crafted features related to Dialog State Tracking (DST). (a) "Chatbot in DST" and "Chatbot providing DST information" (B) "Chatbot starting irrelevant DST" (C) "Chatbot exiting DST".

new conversation session, resulting in a total of 3058 conversation sessions.

3.2.4. Data processing

In order to prepare the dataset for model development, we took several preprocessing steps. First, we removed conversation sessions with more than 10 exchanges to ensure uniform input sequence lengths during model training. We chose this threshold because the majority (83%) of the total 3058 conversation sessions had 10 or fewer exchanges. If we had taken sessions with fewer exchanges, although it results in fewer exchanges needing padding, it would have excluded more sessions for building the model. On the other hand, if we had kept more sessions and loosened the threshold, although we would have kept more data for building the model, more inputs would have needed padding. This excessive padding might have led to data sparsity issues and negatively affected model performance. Sessions with over 25% of the exchanges recorded as "NaN" were also removed, resulting in the elimination of 555 sessions. The remaining 1984 conversation sessions consisting of 8778 conversation exchanges from 1,373 users were used for model development, with 1625 sessions (including 268 abandoning sessions and 61 temporarily leaving sessions) used for training and 359 sessions (including 87 abandoning sessions and 10 temporarily leaving sessions) used for testing.

4. Experiment one: Identifying impactful features from hand-crafted features

4.1. Model development and evaluation

The model for detecting users' abandonment of chatbot was trained using LightGBM (Ke et al., 2017). LightGBM is a gradient boosting decision tree algorithm that incorporates efficient techniques such as gradient-based one-side

sampling (GOSS) and exclusive feature bundling (EFB), which improve the speed of training. It has been shown preferable over other gradient boosting decision tree algorithms like XGBoost and CatBoost due to its high accuracy and efficiency (Al & Daoud, 2019).

The input for LightGBM training consists of conversation sessions, each containing 10 exchanges with 20 hand-crafted features concatenated for each exchange to generate 200 features per session. However, some sessions have fewer than 10 exchanges, and for those sessions, we add padding exchanges at the beginning with -1 as their values shown in Figure 2. We used grid search method to adjust the parameters of LightGBM for training to achieve a better performance and added "is_balanced" to address the issue of imbalanced data, the final parameters were shown in Table 2.

To accurately our model, we employed a five-fold cross-validation technique due to the imbalanced training data. This approach involves dividing the training data into five non-overlapping subsamples. We then trained five separate models, each time using four subsamples for training and one subsample for testing. This process was repeated five times to ensure that each subsample was used for testing exactly once. The models for predicting user abandonment and all achieved a high performance with AUC above 0.83, then we retrained LightGBM on the full training data and tested it on the testing data to obtain the performance of LightGBM.

The model for predicting user abandonment of chatbot demonstrated high performance with an accuracy of 0.92, an F1-score of 0.83, a recall of 0.83, a specificity of 0.95, a precision of 0.83, and AUC of 0.89. These findings indicate that the 20 selected hand-crafted features are useful for predicting instances of chatbot abandonment in the banking chatbot dataset. The model developed to predict instances of user discontinued conversation performed even better,

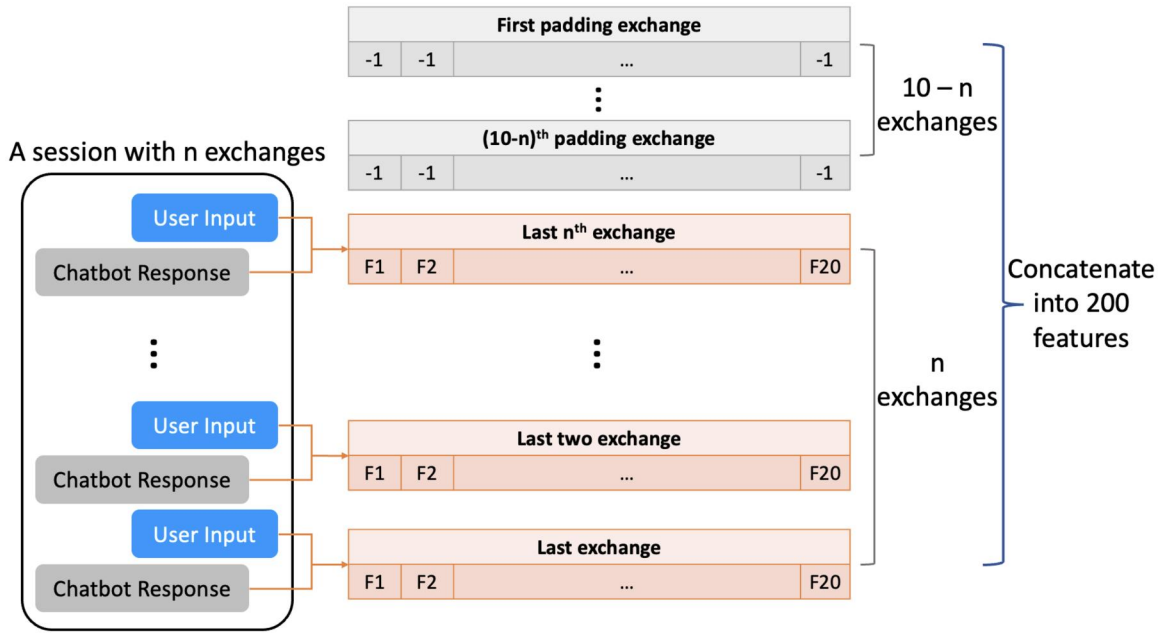


Figure 2. The training input we used in LightGBM. n is the number of the exchanges in a session, which is less than or equal to 10, and each exchange contains 20 hand-crafted features. To ensure uniform input sequence lengths during model training, each session must contain exactly 10 exchanges. When there is less than 10 in a session, padding exchanges are added at the beginning.

Table 2. LightGBM parameters from grid search.

	Abandonment	Discontinued conversation
learning_rate	0.2	0.1
max_depth	5	10
min_child_sample	20	5
num_leaves	31	60
reg_alpha	0.01	0.03

achieving an accuracy of 0.94, an F1-score of 0.88, a recall of 0.85, a specificity of 0.97, a precision of 0.91, and AUC: 0.91. This suggests that the selected hand-crafted features are more predictive of user discontinuing a conversation than abandoning the chatbot.

4.2. Feature analysis

The SHAPley Additive exPlanations (SHAP) values (Lundberg & Lee, 2017) were used to determine the impact of each feature on the model's predictions. This method utilizes Shapely values (Hart, 1989) to show the impact and direction of each feature's effect on the model's predictions, providing insights into the relationship between the features and the predicted outcomes. The impact of each feature on the model prediction is visualized in Figure 3, where the y -axis lists the features in descending order of the average absolute SHAP value (the value in the parenthesis after the feature), and the x -axis represents the influence of the value on the model prediction. Each dot in the chart represents an individual conversation exchange, with color indicating the feature values (ranging from blue for low to red for high), and position on the x -axis showing the positive or negative effect of the feature values on the prediction outcomes.

In Figure 3, it is evident that in both the models for predicting abandonment and discontinued conversation, the most recent exchange held the most weight, with the top

three influential features occurring in the most recent exchange, while earlier exchanges had less of an impact. "User typing" (as opposed to selecting an option) in the most recent conversation exchange was the most impactful feature for both models, with "User typing" in the second most recent conversation exchange also among the top six features. These findings suggest that the way users input information strongly influences the likelihood of them discontinuing the conversation or abandoning the chatbot, possibly due to the fact that typing allows for a wider range of language, phrases, and expressions that the chatbot may struggle to understand accurately.

The features with the second and third highest impact in both models were related to the chatbot's response, specifically "Chatbot non-recognition" (2nd) and "Chatbot providing DST information" (3rd). Interestingly, "Chatbot mis-recognition" had a smaller impact than "Chatbot non-recognition" for both prediction tasks. This suggests that users were more likely to discontinue a conversation or abandon the chatbot when the chatbot indicated that it could not understand their input, rather than when it provided an incorrect output due to misunderstanding their intent. When predicting discontinued conversation, the impact of "Chatbot mis-recognition" was found to be stronger than predicting abandonment. This was demonstrated by the fact that there were no instances labeled as "Chatbot mis-recognition" showing zero or negative SHAP values when predicting discontinued conversation, indicating a strong relationship between this feature and the positive value of the prediction outcome. In contrast, when predicting abandonment, several instances showed zero or negative SHAP values, indicating that some mis-recognition events did not lead to abandonment. This implies that users sometimes left the chatbot only temporarily after the chatbot misunderstood their intent.

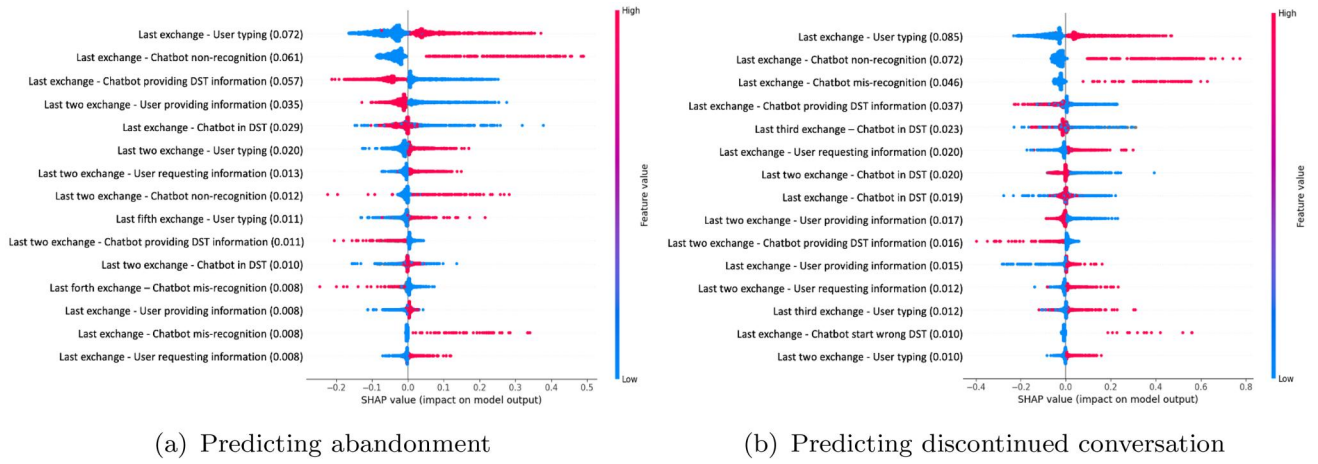


Figure 3. A visualization of the top 15 features identified by SHAP values. The y-axis shows the features in descending order of average absolute SHAP value, while the x-axis represents the degree of influence on the model prediction. Each dot on the graph represents a conversation exchange, with color indicating feature values (ranging from low in blue to high in red). The position on the x-axis represents the positive or negative impact of the feature values on the model's predictions.

Table 3. Comparison of the likelihood of abandonment, temporary leave, and discontinued conversation for Chatbot mis-recognition and non-recognition.

	Abandonment (%)	Temporary leave (%)	Discontinued conversation (%)	Non-leave (%)
Mis- recognition	62	33	95	5
Non- recognition	75	10	85	15

To confirm the previous findings, the study examined the probability of users abandoning versus temporarily leaving the chatbot after mis-recognition and non-recognition events. The results presented in Table 3 revealed that when mis-recognition occurred in the most recent exchange, there was a 62% chance that the user would abandon the chatbot, which was lower than the 75% chance when non-recognition occurred.

However, mis-recognition led users to temporarily leave the chatbot 33% of the time, which was three times more than non-recognition. The findings of the study further support the SHAP interpretation that, when comparing the chatbot's inability to understand the user's input with the chatbot misunderstanding the user's intent, the former more frequently leads to chatbot abandonment, while the latter more frequently results in the user only temporarily leaving the chatbot.

In addition to the impact of chatbot non-recognition and mis-recognition, the feature “User providing information” in the second most recent exchange was found to be impactful in predicting non-abandonment but not in the most recent exchange. Upon further investigation, the absence of this feature in the most recent exchange was likely due to the strong impact of “User typing” overshadowing it. When examining the probability of user abandonment based on how information was provided (typing vs. selecting an option), as Figure 4 shows, we found that the probability of abandonment was 12 times higher when the user provided information by typing in the last exchange compared to selecting an option (12.8%, (22/172) vs. 1% (8/765)). However, the probability decreased to only 3 times higher when the user typed in the second last exchange (8.2%, (20/245) vs. 2.7%, (18/661)). This suggests that the method of

providing information is more influential in the last exchange than in the second last exchange. Possibly because of this, we see the “overshadowing of” “User providing information” by “User typing” in the last exchange, but not the second-to-last exchange.

5. Experiment two: Incorporating textual features into abandonment prediction

In Experiment 1, we demonstrated that it is possible to predict whether a user will abandon a chatbot or discontinue the conversation with the chatbot using hand-crafted features. The SHAP tool was used to identify impactful features that serve as indicators for both leaving events. However, during the process of conducting Experiment 1, we also noticed some issues worth considering. Firstly, hand-crafted features may not capture all the information that reflects users' intention to leave the chatbot. Secondly, it is evidently difficult to achieve real-time prediction using a large number of hand-annotated dialogue features. Therefore, we conducted a second experiment that incorporated both hand-crafted features and input text using BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018). We hoped that this approach would not only capture additional language signals that may be indicative of users abandoning or discontinuing the conversation with the chatbot, beyond what was identified using hand-crafted features alone, but also reduce the need for hand-crafted features and enhance the feasibility of real-time prediction of user abandonment.

5.1. Model development and evaluation

In Experiment 2, we fine-tuned the pre-trained BERT (bert-base-chinese) from Hugging Face (Wolf et al., 2020) to predict both user abandonment and discontinued conversation using both input text and hand-crafted features. However, to adhere to the 512-token input limit of the pre-trained BERT model, only the top four impactful hand-crafted features according to the SHAP results—“User typing,” “Chatbot

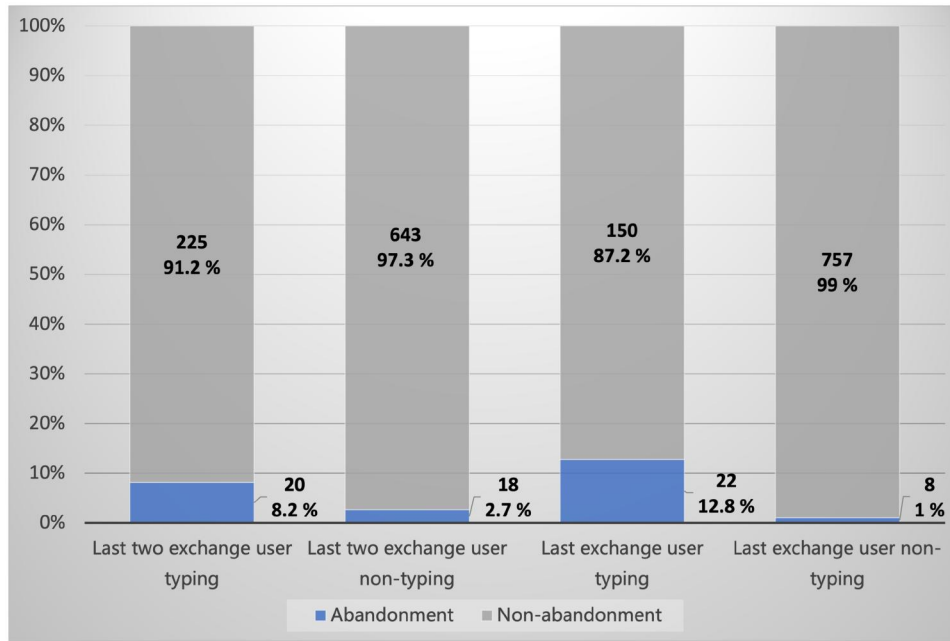


Figure 4. The probability of abandonment when “User providing information” by typing and non-typing. The left two bars are in the last two exchange and the right two bars are in the last exchange.

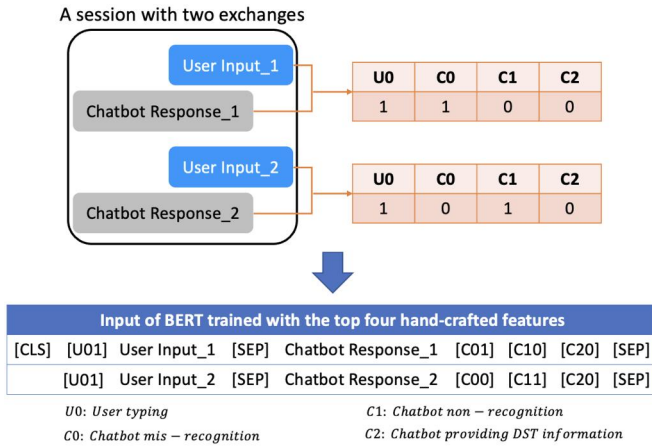


Figure 5. An illustration of the input used for training the pre-trained BERT model, including the top four hand-crafted features identified through SHAP analysis.

non-recognition,” “Chatbot providing DST information,” and “Chatbot mis-recognition”—were added, in addition to “user input” and “chatbot response” as depicted in Figure 5. “User typing” refers to the user’s typed input or selected option. Additionally, to ensure compliance with the 512-token input limit and using the selected four features, any exchange with the four features that exceeds the threshold will be truncated, resulting in a precisely 512-token input length. The input for the model is organized by separating the user input and chatbot response with a *[SEP]* token, combined with the four selected hand-crafted features. Each input is joined by a *[SEP]* token and a *[CLS]* token is added to the beginning of the input.

In this study, a new prediction task was added to the previous tasks of predicting chatbot abandonment and discontinued conversation. The new task involved distinguishing between abandonment and temporary leaving. The aim was

to compare the effectiveness of directly predicting abandonment versus a two-step process of identifying abandonment events among discontinued conversation events. This was done to explore whether the identification of abandonment events among discontinued conversation events would lead to better abandonment prediction, as the LightGBM model performed better in predicting discontinued conversation than abandonment. The results of these four prediction tasks were presented in Table 4 for four model variants, namely BERT trained with the top four hand-crafted features (BERT, Text + Top Features Only), BERT trained with text (BERT, Text), LightGBM using the same four top features (LightGBM, Top Features Only), and LightGBM using the full set of 20 features (LightGBM, Full Features). The metrics of these models were also presented in Table 4. The results of the “distinguishing abandonment from temporary leaving” task, which is the second step in the two-step abandonment prediction, were included to understand why the two-step prediction performed better or worse than the one-step prediction. We also trained the LightGBM model with the same four top features and BERT model with text to assess the impact of incorporating text features from BERT on prediction performance.

All three models, except for BERT (Text), were effective in predicting user abandonment with an AUC of 0.87 or higher in the one-step prediction task, using either hand-crafted features with or without textual features. BERT (Text), which didn’t incorporate hand-crafted features, had a lower AUC of 0.85. Textual features were shown to be better at identifying instances of users abandoning the chatbot, with BERT (Text + Top Features Only) and BERT (Text) achieving the highest Recall scores of 0.85 and 0.82, respectively. BERT (Text + Top Features Only) also outperformed LightGBM (Top Features Only) in Recall (0.85 vs. 0.81). Hand-crafted features were better at improving the precision

Table 4. The performance of the four models, namely, BERT trained with the top four hand-crafted features (BERT, Text + Top Features Only), BERT trained with text (BERT, Text), LightGBM using the same four top features (LightGBM, Top Features Only), and LightGBM using the full set of 20 features (LightGBM, Full Features) in four different prediction tasks: One-step predicting abandonment, predicting discontinued conversation, distinguishing abandonment from temporary leaving, and Two-step predicting abandonment.

Task	Model	Accuracy	F1-score	Recall	Specificity	Precision	AUC
One-Step Predicting Abandonment	LightGBM (Full Features)	0.92	0.83	0.83	0.95	0.83	0.89
	LightGBM (Top Features Only)	0.91	0.81	0.81	0.95	0.82	0.88
	BERT (Text + Top Features Only)	0.90	0.80	0.85	0.91	0.76	0.88
	BERT (Text)	0.86	0.74	0.82	0.88	0.68	0.85
Predicting Discontinued Conversation	LightGBM (Full Features)	0.94	0.88	0.85	0.97	0.91	0.91
	LightGBM (Top Features Only)	0.93	0.85	0.81	0.97	0.90	0.89
	BERT (Text + Top Features Only)	0.94	0.89	0.90	0.96	0.89	0.92
	BERT (Text)	0.93	0.87	0.87	0.95	0.88	0.91
Distinguishing Abandonment from Temporary leaving	LightGBM (Full Features)	0.77	0.87	0.82	0.00	0.88	0.43
	LightGBM (Top Features Only)	0.62	0.76	0.67	0.20	0.88	0.43
	BERT (Text + Top Features Only)	0.83	0.90	0.91	0.10	0.90	0.50
	BERT (Text)	0.80	0.89	0.89	0.10	0.90	0.49
Two-step Predicting Abandonment	LightGBM (Full Features)	0.89	0.75	0.70	0.95	0.81	0.83
	LightGBM (Top Features Only)	0.86	0.64	0.54	0.96	0.80	0.75
	BERT (Text + Top Features Only)	0.89	0.77	0.76	0.93	0.79	0.85
	BERT (Text)	0.89	0.78	0.77	0.93	0.79	0.85

of the prediction, with LightGBM (Full Features) and LightGBM (Top Features Only) achieving the first and second-highest Precision of 0.83 and 0.82, respectively. Furthermore, including more hand-crafted features improved precision, with BERT (Text + Top Features Only) achieving a Precision of 0.76, significantly higher than BERT (Text) with a Precision of 0.68. Interestingly, inclusion of non-top hand-crafted features into LightGBM models improved precision but decreased recall, while inclusion of top hand-crafted features into BERT improved recall and precision. As a result, BERT (Text + Top Features Only) outperformed BERT (Text) in all performance metrics, showing the great benefits of including top hand-crafted features in BERT for predicting chatbot abandonment.

The two-step prediction approach was not as effective for hand-crafted features as it was for textual features. In this prediction task, LightGBM models, which did not consider textual features, had significantly lower Recall, whereas the two BERT models did not experience as much of a drop in recall as the two LightGBM models, and as a result, they outperformed the LightGBM models in both F1-scores and AUC. This result suggests that models that leverage textual features have much better performance in capturing abandonment in discontinued conversations compared to those using hand-crafted features. This observation is also supported by the nearly equivalent performance between BERT (Text) and BERT (Text + Top Features Only) in this prediction task, with the former even achieving slightly better Recall. This suggests that hand-crafted features did not provide much information that helps identify chatbot abandonment when differentiating it from temporary leaving. As a result, all models that used hand-crafted features, except BERT (Text), led to both lower AUC and F1-scores in the two-step prediction task than in the one-step prediction task.

5.2. Leaving signals suggested by LIME

To better understand the textual features that contribute to the BERT model's predictions of whether a user will

abandon a chatbot service, we employed LIME (Local Interpretable Model-Agnostic Explanations), a widely used XAI technique for interpreting complex models' predictions (Bagora et al., 2022; Hernandez Urbano et al., 2021; Kim et al., 2022). LIME (Ribeiro et al., 2016) generates a visualization of words that play a significant role in the prediction outcome of abandonment, allowing us to identify possible leaving signals conveyed through text. We observed that two chatbot responses are frequently marked by LIME as a significant contributor to users abandoning the chatbot: "Sorry [the chatbot's name] did not understand your question very clearly, could you please try asking it in a different way?" and "What would you like to do?" We note that these phrases are specific to the chatbot we analyzed and were originally in Mandarin; but we report this results because we assume phrases that convey a similar meaning from other chatbots may lead to a similar result.

Figure 6 depicts an instance where the chatbot used the phrase in response to the user's input. The dark orange highlighting of the phrase in the figure indicates that it played the most significant role in the model's prediction of the user abandoning the chatbot, even though the four hand-crafted features were present in the input. However, we found that merely using this phrase does not always make it a significant predictor of leaving. We noted that whether the phrase was identified as a strong predictor depended on whether the chatbot continued the conversation after using it. Specifically, we found that if the user left the chatbot after the phrase was used, the model always identified it as the most important sign.

On the other hand, Figure 7 illustrates an example where the chatbot continued the conversation after the phrase was used, because the user managed to ask a question the chatbot understood. In this case, the model considered a follow-up question (highlighted in dark blue) to be a more significant sign of the user continuing the conversation. The examples presented show that taking into account the language context of the chatbot's output is essential for accurately identifying signs of users leaving, and the fact that the highlighted phrases are more important predictors than the

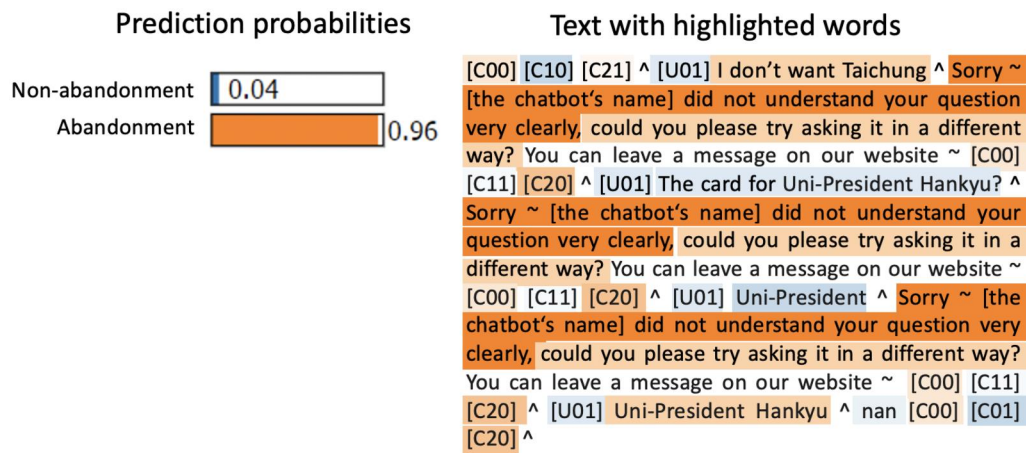


Figure 6. English translation of the LIME result focusing on “Sorry [the chatbot’s name] did not understand your question very clearly, could you please try asking it in a different way?” is considered highly likely to cause chatbot abandonment.

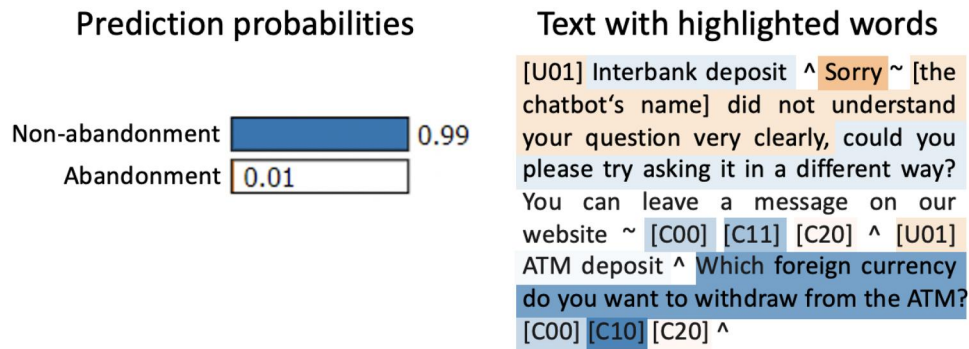


Figure 7. English translation of the LIME result focusing on “Sorry [the chatbot’s name] did not understand your question very clearly, could you please try asking it in a different way?”, but it is not as predictive for chatbot abandonment due to the presence of the user’s next question being answered.

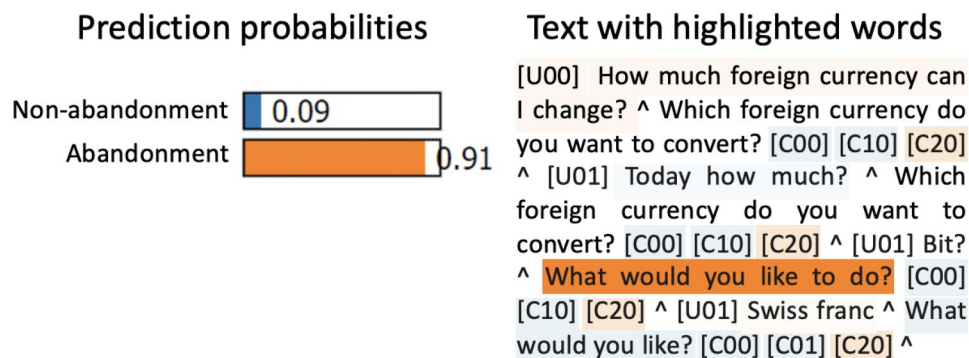


Figure 8. English translation of the LIME result focusing on “What would you like to do?” when it is presented in the middle of an ongoing conversation.

four hand-crafted features reinforces the idea that textual signals are valuable for predicting user abandonment.

The phrase “What would you like to do?” was also frequently identified by LIME as a significant contributor to users abandoning the chatbot. This phrase was often used by the chatbot to reset a conversation, even when the user had already expressed an explicit request, as shown in Figure 8. While asking a new question may seem like a reasonable way to move past a conversation breakdown and continue the conversation, our results show that presenting this question in the middle of an ongoing conversation, especially when the user has already

clearly expressed their intent, tended to contribute to users leaving the chatbot. Users who still wanted to continue in the same conversation found this phrase unhelpful, as it assumed that they wanted to keep using the chatbot without providing guidance on how to proceed. As a result, we observed that when users encountered this phrase, they usually repeated their previous statement, attempted to continue their previous line of inquiry, and ultimately abandoned the chatbot. This indicates that this phrase not only failed to help the conversation progress but also disrupted the user’s ongoing attempts, leading to a poor user experience.

6. Discussion

6.1. Feasibility of predicting chatbot abandonment

The study aimed to evaluate the effectiveness of different methods for predicting user abandonment in chatbot interactions. The results show that all three models that leveraged hand-crafted features achieved high F1-scores (at least 0.80) and AUC (at least 0.87), indicating that predicting user abandonment using conversation logs and hand-crafted features is a viable and reliable approach.

However, textual features were found to be more effective than hand-crafted features in capturing chatbot abandonment events, as evidenced by the higher Recall scores of the BERT models. Textual features were particularly useful for differentiating between users abandoning the chatbot and users who only temporarily leave. This was supported by the two-step prediction task, where the LightGBM models had significantly lower Recall than they had in the one-step prediction, whereas the two BERT models did not experience as much of a drop in Recall as the two LightGBM models. Hand-crafted features were still good at differentiating between continuous and discontinued conversation events but were not as sensitive as textual features in further differentiating the two types of leaving events. This may be due to the nuanced differences between the two types of leaving events that require more fine-grained features. As a result, in the two-step prediction approach, BERT (Text) outperformed the two LightGBM models in both F1-scores, AUC, and Recall, making it a favorable choice for predicting chatbot abandonment compared to other models, despite being less favorable in the one-step prediction.

Based on our findings, a chatbot service provider can choose to use a pre-trained BERT model with only textual features and the two-step prediction approach to reduce the need for manually labeling conversation logs while still achieving equivalent performance metrics in predicting chatbot abandonment. However, if the provider wants to capture as many abandonment events as possible, we recommend using a pre-trained BERT model with impactful hand-crafted features, even though it requires manual effort to create them. This model can achieve a Recall of up to 0.85 and an AUC of up to 0.88. However, having many false positives can increase the need for human intervention, so providers may opt for a model with higher precision and fewer false alarms by using hand-crafted features. The top impactful features such as non-recognition, user typing, chatbot providing information in DST, can be determined and added automatically by a chatbot system, as long as the chatbot constantly tracks the user's states. Additionally, XAI tools like SHAP can be used to analyze the effectiveness of these features in predicting user abandonment.

The decision to use hand-crafted features or BERT-based models should take into account additional factors, such as whether discontinued conversations should also be captured and the need for real-time detection. Although some hand-crafted features can be automatically generated, they may introduce latency compared to a BERT model that uses only textual features. Ultimately, the choice should depend on the service provider's specific needs and the trade-off between the cost of creating hand-crafted features and the risk of false alarms.

6.2. Leaving signals of chatbot abandonment

From the two experiments, we obtained some leaving signals from LightGBM using full set of hand-crafted features and BERT trained with the top four hand-crafted features. We found that these leaving signals can be divided into two types (1) The way the user interacts with the chatbot and (2) Chatbot's response. We will discuss based on these types.

6.2.1. The way user interact with chatbot

Our analysis using LightGBM demonstrated that the manner in which users interacted with the chatbot greatly impacted the likelihood of chatbot abandonment. That is, While earlier research has shown that free typing reduces both the pragmatic and hedonic user experience compared to button interaction (Haugeland et al., 2022), our study demonstrates that free typing is also much more likely to lead to chatbot abandonment than pre-set response selection. Notably, our findings showed that when users selected a pre-set response in the last exchange, the probability of chatbot abandonment was only 1%, which was significantly lower than the 12.8% probability of abandonment when users provided free-form messages by typing. These findings are consistent with prior studies suggesting that providing pre-set options to users can reduce the occurrence of conversation breakdowns (Ashktorab et al., 2019; Jain et al., 2018; Li et al., 2020; Yeh et al., 2022). It is likely because the use of a wider range of language, phrases, and expressions in user input may make the chatbot misunderstand or make it difficult for the chatbot to understand the user's intent. Additionally, user inputs may be outside the scope of the chatbot or contain errors such as misspellings or unfinished sentences. These results suggest that chatbot designers may benefit from allowing users to select pre-set responses or providing constraints and guidelines for user inputs to improve the overall conversation experience and reduce the likelihood of chatbot abandonment.

6.2.2. Chatbot's response

Our analysis using both SHAP and LIME suggests that the feature "Chatbot non-recognition" strongly predicts both discontinued conversation and abandonment. When the chatbot directly expresses its inability to recognize the user's intent, the user may perceive the chatbot as incompetent and unlikely to serve their needs, potentially leading to abandonment rather than just temporary leaving. In contrast, when "Chatbot mis-recognition" occurred, there was a lower probability of abandonment and a higher likelihood that the leaving was only temporary compared to non-recognition. This may be because users appreciate the chatbot's effort to understand their intent, even if it gives a wrong answer, leading them to perceive mis-recognition as a minor issue rather than a serious one as non-recognition.

In addition to the strong predictors of chatbot abandonment, the LIME analysis reveals that a simple question from the chatbot asking what the user would like to do in the middle of an ongoing conversation is also a strong signal of

user abandonment. Although this response from the chatbot has not been considered a conversation non-progress in previous research (Li et al., 2020), our results show that it is perceived by the model as strongly contributing to the outcome of user abandonment. This may be because users perceive the chatbot as disregarding their current attempt and restarting the conversation, which fails to provide helpful information or address their intent, which was recommended in prior studies as helpful for chatbot users (Ashktorab et al., 2019). This response may also suggest the chatbot's incompetence in recognizing the user's intent.

These findings suggest that chatbot practitioners should avoid expressing incompetence, inability, or ignoring the user's ongoing attempt, and instead, present possible intents to the user to avoid direct user conclusions that the chatbot is incapable, which could potentially harm the reputation of the business that provides the chatbot. When the chatbot is uncertain, notifying the user that the conversation is being transferred to customer support could also prevent the user from abandoning the chatbot. However, further studies are needed to validate the effectiveness of this strategy.

6.3. Research limitation

There are several important limitations to the current study that must be considered when interpreting the results. First, the study was conducted using a conversation log from a specific banking chatbot, and the hand-crafted features used were created by the authors of Li et al. (2020). The extent to which the results of this study can be generalized to predict chatbot abandonment in other domains is unclear, and it is possible that other impactful features were not explored. Therefore, it is uncertain to what extent the results can be generalized to other chatbots or domains, and it is possible that other impactful features were not included. As the research only explored a limited number of hand-crafted features, it cannot be claimed that it considers an exhaustive list of such features. Future research could investigate other signs of chatbot abandonment that were not explored in this study.

Secondly, the study only included sessions with a limited number of exchanges and also restricted the number of hand-crafted features used in each exchange. This may limit the performance of the models. When adding hand-crafted features in training, the number of text in each exchange will be relatively reduced, which may lower the performance. Thirdly, while the study identified leaving signals through XAI tools, such as SHAP and LIME, the actual reasons for users abandoning the chatbot were not known. Therefore, observational studies in combination with debriefing interviews would be necessary to gain insight into the users' intentions and reasons for leaving the chatbot.

Finally, the study using hand-crafted features may not achieve real-time prediction, we suggest future research to focus on using text models, such as BERT, to convert text into these features first and then make predictions. This approach can achieve real-time prediction and outperform simply using text to predict.

7. Conclusion

As more businesses turn to task-oriented chatbots to deliver their services, it is crucial to anticipate instances where users may abandon the chatbot. This paper demonstrated the feasibility of predicting user abandonment based on an annotated conversation log involving 1373 users. SHAP and LIME were employed to identify impactful features that contribute to chatbot abandonment, including user typing, chatbot inability to recognize intent, and the chatbot asking what users want to do during an ongoing conversation. Our findings suggest that chatbot designers should prioritize recognizing the user's intent, as a failure to do so is more likely to lead to abandonment. Additionally, we identified a third type of conversation breakdown where the chatbot arbitrarily restarts the ongoing conversation, which is often neglected but highly predictive of chatbot abandonment. In light of our results, we recommend that chatbot designers increase the occurrence of presenting pre-set options or constraints for user inputs and providing possible intents to the user. Furthermore, they should avoid expressions of inability, incompetence, or ignoring the user's current attempt to potentially prevent abandonment.

Acknowledgement

We sincerely thank the bank institution that provided us with the conversation log for our research. This research was supported by the National Science and Technology Council, Taiwan, R.O.C (1092218-E-009 -016, 1102222-E-A49-008-MY3).

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Chieh Hsu  <http://orcid.org/0009-0001-0143-1246>
 Hsin-Chien Tung  <http://orcid.org/0009-0004-0235-2138>
 Hong-Han Shuai  <http://orcid.org/0000-0003-2216-077X>
 Yung-Ju Chang  <http://orcid.org/0000-0001-6956-3459>

References

- Acikgoz, F., & Vega, R. P. (2022). The role of privacy cynicism in consumer habits with voice assistants: A technology acceptance model perspective. *International Journal of Human-Computer Interaction*, 38(12), 1138–1152. <https://doi.org/10.1080/10447318.2021.1987677>
- Al Daoud, E. (2019). Comparison between xgboost, lightgbm and catboost using a home credit dataset. *International Journal of Computer and Information Engineering*, 13(1), 6–10. <https://doi.org/10.5281/zenodo.3607805>
- Amer, E., Hazem, A., Farouk, O., Louca, A., Mohamed, Y., & Ashraf, M. (2021). A proposed chatbot framework for covid-19. In *2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)* (pp. 263–268). IEEE. <https://doi.org/10.1109/MIUCC52538.2021.9447652>
- Amershi, S., Weld, D., Vorvoreanu, M., Fournay, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., & Teevan, J. (2019). Guidelines for human-AI interaction. In *Proceedings of the 2019 Chi Conference on Human Factors in Computing Systems* (pp. 1–13). ACM. <https://doi.org/10.1145/3290605.3300233>
- Ashktorab, Z., Jain, M., Liao, Q., & Weisz, J. D. (2019). Resilient chatbots: Repair strategy preferences for conversational breakdowns. In

- Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM.
- Athreya, R. G., Ngonga Ngomo, A.-C., & Usbeck, R. (2018). Enhancing community interactions with data-driven chatbots—the DBpedia chatbot. In *Companion Proceedings of the Web Conference 2018* (pp. 143–146). ACM.
- Atif, M., Hassan, M. K., Rabbani, M. R., & Khan, S. (2021). Islamic fintech: The digital transformation bringing sustainability to Islamic finance. In *Covid-19 and Islamic social finance* (pp. 91–103). Routledge.
- Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One*, 10(7), e0130140. <https://doi.org/10.1371/journal.pone.0130140>
- Bagana, B. D., Irsad, M., & Santoso, I. H. (2021). Artificial intelligence as a human substitution? customer's perception of the conversational user interface in banking industry based on Utaut concept. *Review of Management and Entrepreneurship*, 5(1), 33–44. <https://doi.org/10.37715/rme.v5i1.1632>
- Bagora, A., Shrestha, K., Maurya, K., & Desarkar, M. S. (2022). Hostility detection in online hindi-english code-mixed conversations. In *14th ACM Web Science Conference 2022* (pp. 390–400). ACM. <https://doi.org/10.1145/3501247.3531579>
- Bhaskaran, S. (2020). How financial service providers use AI-based chatbots to improve customer experience. <https://www.twilio.com/hub/how-financial-services-providers-are-using-ai-powered-chatbots-level-their-customer-experience>
- Bouhia, M., Rajaobelina, L., PromTep, S., Arcand, M., & Ricard, L. (2022). Drivers of privacy concerns when interacting with a chatbot in a customer service encounter. *International Journal of Bank Marketing*, 40(6), 1159–1181. <https://doi.org/10.1108/IJBM-09-2021-0442>
- Brandtzaeg, P. B., & Følstad, A. (2017). Why people use chatbots. In *Internet Science: 4th International Conference, INSCI 2017, Thessaloniki, Greece, November 22–24, 2017, Proceedings 4* (pp. 377–392). Springer.
- Buhalis, D., & Cheng, E. S. Y. (2020). Exploring the use of chatbots in hotels: Technology providers' perspective. In *Information and Communication Technologies in Tourism 2020: Proceedings of the International Conference in Surrey, United Kingdom, January 8–10, 2020* (pp. 231–242). Springer.
- Catledge, L. D., & Pitkow, J. (1995). Characterizing browsing strategies in the world-wide web. *Computer Networks and ISDN Systems*, 27(6), 1065–1073. [https://doi.org/10.1016/0169-7552\(95\)00043-7](https://doi.org/10.1016/0169-7552(95)00043-7)
- Chaves, A. P., & Gerosa, M. A. (2019). How should my chatbot interact? A survey on human-chatbot interaction design. *International Journal of Human-Computer Interaction*, 37(8), 729–758. <https://doi.org/10.1080/10447318.2020.1841438>
- Choi, I. H., Kim, Y. S., & Lee, C. K. (2020). A study of the classification of it jobs using LSTM and lime. In *The 9th International Conference on Smart Media and Applications* (pp. 248–252). ACM. <https://doi.org/10.1145/3426020.3426083>
- da Silva, T. H., Furtado, V., Furtado, E., Mendes, M., Almeida, V., & Sales, L. (2022). How do illiterate people interact with an intelligent voice assistant? *International Journal of Human-Computer Interaction*, 1–19. <https://doi.org/10.1080/10447318.2022.2121219>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Ebermann, C., Selisky, M., & Weibelzahl, S. (2023). Explainable AI: The effect of contradictory decisions and explanations on users' acceptance of AI systems. *International Journal of Human-Computer Interaction*, 39(9), 1807–1826. <https://doi.org/10.1080/10447318.2022.2126812>
- Engelhardt, S., Hansson, E., & Leite, I. (2017). Better faulty than sorry: Investigating social recovery strategies to minimize the impact of failure in human-robot interaction. In *WCII/HAI@IVA*. Emerald Publishing Limited.
- Eren, B. A. (2021). Determinants of customer satisfaction in chatbot use: Evidence from a banking application in turkey. *International Journal of Bank Marketing*, 39(2), 294–311. <https://doi.org/10.1108/IJBM-02-2020-0056>
- Farah, J. C., Spaenlehauer, B., Ingram, S., & Gillet, D. (2022). A blueprint for integrating task-oriented conversational agents in education. In *Proceedings of the 4th Conference on Conversational User Interfaces* (pp. 1–8). ACM. <https://doi.org/10.1145/3543829.3544525>
- Følstad, A., & Bjerkreim-Hanssen, N. (2023). User interactions with a municipality chatbot—lessons learnt from dialogue analysis. *International Journal of Human-Computer Interaction*, 1–14. <https://doi.org/10.1080/10447318.2023.2238355>
- Følstad, A., Skjuve, M., & Brandtzaeg, P. B. (2019). Different chatbots for different purposes: Towards a typology of chatbots to understand interaction design. In *Internet Science: INSCI 2018 International Workshops, St. Petersburg, Russia, October 24–26, 2018, revised selected papers 5* (pp. 145–156). Springer.
- Grudin, J., & Jacques, R. (2019). Chatbots, humbots, and the quest for artificial general intelligence. In *Proceedings of the 2019 Chi Conference on Human Factors in Computing Systems* (pp. 1–11). ACM. <https://doi.org/10.1145/3290605.3300439>
- Handoyo, E., Arfan, M., Soetrisno, Y. A. A., Somantri, M., Sofwan, A., & Sinuraya, E. W. (2018). Ticketing chatbot service using serverless NLP technology. In *2018 5th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)* (pp. 325–330). IEEE. <https://doi.org/10.1109/ICITACEE.2018.8576921>
- Hart, S. (1989). *Shapley value*. Springer.
- Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer.
- Haugeland, I. K. F., Følstad, A., Taylor, C., & Bjørkli, C. A. (2022). Understanding the user experience of customer service chatbots: An experimental study of chatbot interaction design. *International Journal of Human-Computer Studies*, 161, 102788. <https://doi.org/10.1016/j.ijhcs.2022.102788>
- Henderson, M., Thomson, B., & Williams, J. D. (2014). *The third dialog state tracking challenge*. In *2014 IEEE Spoken Language Technology Workshop (SLT)* (pp. 324–329). IEEE. <https://doi.org/10.1109/SLT.2014.7078595>
- Hernandez Urbano, R., Jr., Uy Ajero, J., Legaspi Angeles, A., Hacar Quintos, M. N., Regalado Imperial, J. M., & Llabanes Rodriguez, R. (2021). A Bert-based hate speech classifier from transcribed online short-form videos. In *2021 5th International Conference on e-Society, e-Education and e-Technology* (pp. 186–192). ACM. <https://doi.org/10.1145/3485768.3485806>
- Hong, D., & Cho, C.-H. (2022). Factors affecting innovation resistance of smartphone AI voice assistants. *International Journal of Human-Computer Interaction*, 39(13), 1–16. <https://doi.org/10.1080/10447318.2022.2080899>
- Hussain, S., Ameri Sianaki, O., & Ababneh, N. (2019). A survey on conversational agents/chatbots classification and design techniques. In *Web, artificial intelligence and network applications: Proceedings of the workshops of the 33rd international conference on advanced information networking and applications (WAINA-2019)* 33 (pp. 946–956). Springer.
- Jain, M., Kumar, P., Kota, R., & Patel, S. (2018). Evaluating and informing the design of chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference*. ACM. <https://doi.org/10.1145/3196709.3196735>
- Jones, R., & Klinkner, K. (2008). Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs. In *CIKM'08*. ACM.
- Kane, D. A. (2016). The role of chatbots in teaching and learning. *E-Learning and the Academic Library: Essays on Innovative Initiatives*, 131, 131–147.
- Kannagi, L., Ramya, C., Shreya, R., & Sowmiya, R. (2018). Virtual conversational assistant: 'The farmbot'. *International Journal of Engineering Technology Science and Research*, 5(3), 520–527.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 3149–3157.
- Khan, S., & Rabbani, M. R. (2021). Artificial intelligence and NLP-based chatbot for Islamic banking and finance. *International Journal of Information Retrieval Research*, 11(3), 65–77. <https://doi.org/10.4018/IJIRR.2021070105>

- Kim, J., Lee, B., & Sohn, K.-A. (2022). Why is it hate speech? Masked rationale prediction for explainable hate speech detection. In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 6644–6655).
- Langevin, R., Lordon, R. J., Avrahami, T., Cowan, B. R., Hirsch, T., & Hsieh, G. (2021). Heuristic evaluation of conversational agents. In *Proceedings of the 2021 Chi Conference on Human Factors in Computing Systems* (pp. 1–15). ACM. <https://doi.org/10.1145/3411764.3445312>
- Lee, D., Oh, K., & Choi, H.-J. (2017). The chatbot feels you - a counseling service using emotional response generation. In *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)* (pp. 437–440). IEEE.
- Lee, M. K., Kiesler, S., Forlizzi, J., Srinivasa, S., & Rybski, P. (2010). Gracefully mitigating breakdowns in robotic services. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 203–210). ACM.
- Lee, S., Oh, J., & Moon, W.-K. (2022). Adopting voice assistants in online shopping: Examining the role of social presence, performance risk, and machine heuristic. *International Journal of Human-Computer Interaction*, 39(14), 1–15. <https://doi.org/10.1080/10447318.2022.2089813>
- Li, C.-H., Yeh, S.-F., Chang, T.-J., Tsai, M.-H., Chen, K., & Chang, Y.-J. (2020). A conversation analysis of non-progress and coping strategies with a banking task-oriented chatbot. In *Proceedings of the 2020 Chi Conference on Human Factors in Computing Systems* (pp. 1–12). ACM. <https://doi.org/10.1145/3313831.3376209>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4768–4777. <https://dl.acm.org/doi/10.5555/3295222.3295230>
- Mao, Q., Liu, G., Chen, Z., Guo, J., & Liu, P. (2022). Loan prepayment prediction based on SVM-RFE and XGBoost models. In *Proceedings of the International Conference on Information Economy, Data Modeling and Cloud Computing, ICIDC 2022, 17–19 June 2022, Qingdao, China*. EAI.
- Mulyono, J. A., & Sfenrianto, S. (2022). Evaluation of customer satisfaction on indonesian banking chatbot services during the covid-19 pandemic. *CommIT (Communication and Information Technology) Journal*, 16(1), 69–85. <https://doi.org/10.21512/commit.v16i1.7813>
- Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. *Journal of Chemometrics*, 18(6), 275–285. <https://doi.org/10.1002/cem.873>
- Nauta, M., Bucur, D., & Seifert, C. (2019). Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1), 312–340. <https://doi.org/10.3390/make1010019>
- Nguyen, Q. N., Ta, A., & Prybutok, V. (2019). An integrated model of voice-user interface continuance intention: The gender effect. *International Journal of Human-Computer Interaction*, 35(15), 1362–1377. <https://doi.org/10.1080/10447318.2018.1525023>
- Nielsen, J. (1995). How to conduct a heuristic evaluation. *Nielsen Norman Group*, 1(1), 8.
- Okuda, T., & Shoda, S. (2018). AI-based chatbot service for financial industry. *Fujitsu Scientific and Technical Journal*, 54(2), 4–8.
- Parsa, A. B., Movahedi, A., Taghipour, H., Derrible, S., & Mohammadian, A. K. (2020). Toward safer highways, application of xgboost and shap for real-time accident detection and feature analysis. *Accident; Analysis and Prevention*, 136, 105405. <https://doi.org/10.1016/j.aap.2019.105405>
- Prasetyo, P. K., Achananuparp, P., & Lim, E.-P. (2020). Foodbot: A goal-oriented just-in-time healthy eating interventions chatbot. In *Proceedings of the 14th EAI International Conference on Pervasive Computing Technologies for Healthcare* (pp. 436–439). ACM.
- Rapp, A., Curti, L., & Boldi, A. (2021). The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies*, 151, 102630. <https://doi.org/10.1016/j.ijhcs.2021.102630>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?” Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM.
- Rigatti, S. J. (2017). Random forest. *Journal of Insurance Medicine (New York, N.Y.)*, 47(1), 31–39. <https://doi.org/10.17849/insm-47-01-31-39.1>
- Segura, C., Palau, À., Luque, J., Costa-Jussà, M. R., & Banchs, R. E. (2019). Chatbol, a chatbot for the Spanish “La Liga”. In *9th International Workshop on Spoken Dialogue System Technology* (pp. 319–330). Springer.
- Silva, G. R. S., & Canedo, E. D. (2022). Towards user-centric guidelines for chatbot conversational design. *International Journal of Human-Computer Interaction*, 1–23. <https://doi.org/10.1080/10447318.2022.2118244>
- Smutny, P., & Schreiberova, P. (2020). Chatbots for learning: A review of educational chatbots for the facebook messenger. *Computers & Education*, 151, 103862. <https://doi.org/10.1016/j.compedu.2020.103862>
- Tenemaza, M., Luján-Mora, S., de Antonio, A., Ramírez, J., & Zarabia, O. (2020). Ekybot: framework proposal for chatbot in financial enterprises. In *Intelligent human systems integration 2020: Proceedings of the 3rd international conference on intelligent human systems integration (IHSI 2020): Integrating people and intelligent systems, February 19–21, 2020* (pp. 254–259). Springer.
- Uchida, S. (2013). Image processing and recognition for biological images. *Development, Growth & Differentiation*, 55(4), 523–549. <https://doi.org/10.1111/dgd.12054>
- van Heel, M., Harauz, G., Orlova, E. V., Schmidt, R., & Schatz, M. (1996). A new generation of the imagic image processing system. *Journal of Structural Biology*, 116(1), 17–24. <https://doi.org/10.1006/jsbi.1996.0004>
- Vassos, S., Malliaraki, E., Falco, F. D., Di Maggio, J., Massimetti, M., Nocentini, M. G., & Testa, A. (2016). Art-bots: Toward chat-based conversational experiences in museums. In *Interactive Storytelling: 9th International Conference on Interactive Digital Storytelling, ICIDS 2016, Los Angeles, CA, USA, November 15–18, 2016, Proceedings 9* (pp. 433–437). Springer.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *ArXiv, abs/1706.03762*.
- Viswanath Prakash, A., & Das, S. (2020). Would you trust a bot for healthcare advice? An empirical investigation. In *PACIS 2020 Proceedings* (Vol. 62). <https://aisel.aisnet.org/pacis2020/62>
- Weisz, J. D., Jain, M., Joshi, N. N., Johnson, J., & Lange, I. (2019). Bigbluebot: Teaching strategies for successful human-agent interactions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM.
- Williams, K., & Zitouni, I. (2017). Does that mean you’re happy? RNN-based modeling of user interaction sequences to detect good abandonment. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 727–736). ACM.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Brew, J. (2020). Transformers: State-of-the-art natural language processing (pp. 38–45). Association for Computational Linguistics. <https://www.aclweb.org/anthology/2020.emnlp-demos.6>
- Wu, D., & Zhang, S. (2020). Prediction of good abandonment behavior in mobile search. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval* (pp. 407–411). ACM. <https://doi.org/10.1145/3343413.3378007>
- Xue, H., Yang, Q., & Chen, S. (2009). SVM: Support vector machines. In *The top ten algorithms in data mining* (pp. 51–74). Chapman and Hall/CRC.
- Yarosh, S., Thompson, S., Watson, K., Chase, A., Senthilkumar, A., Yuan, Y., & Brush, A. B. (2018). Children asking questions: Speech interface reformulations and personification preferences. In *Proceedings of the 17th ACM Conference on Interaction Design and Children* (pp. 300–312). ACM.
- Yeh, S.-F., Wu, M.-H., Chen, T.-Y., Lin, Y.-C., Chang, X., Chiang, Y.-H., & Chang, Y.-J. (2022). How to guide task-oriented chatbot users, and when: A mixed-methods study of combinations of chatbot guidance types and timings. In *Proceedings of the 2022 Chi Conference*

on *Human Factors in Computing Systems* (pp. 1–16). ACM. <https://doi.org/10.1145/3491102.3501941>

Zhang, Q., Wu, Y., & Zhu, S. (2018). Interpretable convolutional neural networks. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8827–8836). ACM.

Zhou, H., Young, T., Huang, M., Zhao, H., Xu, J., & Zhu, X. (2018). Commonsense knowledge aware conversation generation with graph attention. In IJCAI (pp. 4623–4629). AAAI Press.

About the authors

Chieh Hsu received her BS degree from the Department of Computer Science and Engineering, National Chung Hsing University (NCHU), Taiwan. She is now a graduate student of the Department of Multimedia Engineering in the National Yang Ming Chiao Tung University (NYCU). Her research focuses on human computer interaction.

Hsin-Chien Tung received his BS degree from the Department of Computer Science and Engineering, National Chung Hsing University (NCHU), Taiwan. He is now a graduate student of the Department of Computer Science and Engineering in the National Yang Ming Chiao Tung University (NYCU). His research focuses on human computer interaction.

Hong-Han Shuai received the BS, MS and PhD degree from the Department of Electrical Engineering, National Taiwan University (NTU), Taiwan. He is now an associate professor in National Yang Ming Chiao Tung University (NYCU), Taiwan. His research interests are multimedia processing, machine learning, social network analysis, and data mining.

Yung-Ju Chang is an associate professor at the Department of Computer Science at NYCU. He received his MS and PhD degree in Information Science from the University of Michigan. His research interest is human computer interaction (HCI), with a focus on attention research in computer-mediated communication.