Large-scale Hybrid Approach for Predicting User Satisfaction with Conversational Agents

Dookun Park Hao Yuan Dongmin Kim Yinglei Zhang Spyros Matsoukas
Young-Bum Kim Ruhi Sarikaya Chenlei Guo Yuan Ling
Kevin Quinn Tuan-Hung Pham Benjamin Yao Sungjin Lee
Alexa AI, Amazon
300 Pine St., Seattle, Washington, United States
{dkpark, yuanha, kdongmin, yingleiz, matsouka, youngbum, rsarikay, guochenl, yualing, kevquinn, hupha, benjamy, sungjinl}@amazon.com

Abstract

Measuring user satisfaction level is a challenging task, and it is a critical component in developing large-scale conversational agent systems serving real users. A widely used approach to tackle this is to collect human annotation data and use them for evaluation or modeling. Human annotation based approaches are easier to control, but they are hard to scale. A novel alternative approach is to collect user's direct feedback via a feedback elicitation system embedded to the conversational agent system and to use the collected user feedback to train a machine-learned model for generalization. User feedback is the best proxy for user satisfaction, but it is not available for some ineligible intents and certain situations. Also, asking too much feedback can hurt user experience. Thus, these two types of approaches are complementary to each other. In this work, we tackle a user satisfaction assessment problem with a hybrid approach that fuses explicit user feedback and user satisfaction predictions inferred by two machine-learned models, one trained on user feedback data and the other human annotation data. With this approach, both human annotators and users are involved in the model development process loop as critical roles. The hybrid approach is based on a waterfall [1] policy, and the experimental results with Amazon Alexa's large-scale data sets show significant improvements in inferring user satisfaction. A detailed hybrid architecture, an in-depth analysis on user feedback data, and an algorithm that generates data sets to properly simulate the live traffic are presented in this paper.

1 Introduction

Developing intelligent conversational agents (1) is a topic of great interest in Artificial Intelligence, and there are already several large-scale conversational agents on the market, such as Alexa, Google Assistant, Siri, and Cortana, that have significant numbers of users and usages (2; 3). Nowadays, conversational agents employ a diverse set of metrics to capture different aspects of business and user experience. For example, there are topline business metrics to track and drive success of the business, including monthly active user, dialog count per user, downstream impact from highly valued action and negatively valued action. However, these metrics are normally not sensitive enough to detect changes in a timely fashion, requiring long experimentation times, and hence will result in slow experiment turnaround cycle. On the other hand, there are user experience metrics that capture unhandled user requests or dissatisfying user experiences due to system errors, incomplete service coverage, or poor response quality. User experience metrics generally have positive correlation with

[1] A strategy with linear sequential structure, where each phase depends on the outcome of the previous phase.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

business metrics (4; 5; 6) and tend to be fast moving and more sensitive than business metrics, and hence are suitable for experimentation to make data-driven decisions. Fine-grained user experience metrics are also a key for providing actionable insights to conversational skill developers by correlating metric shifts with interpretable factors such as user intents and slots.

In industry, manual annotation has been widely adopted to assess user satisfaction. For a user utterance, a conversational agent makes a decision that maps the utterance to the best skill/outlet that can handle the user request using contextual signals and machine learning models (7; 8). Given a pair of the user utterance and the agent response, annotators return a user satisfaction score based on annotation guidelines and cross-annotator-calibrations. However, due to its offline nature and resource limitation, human annotation is ill-suited for online monitoring and experimentation or for providing actionable insights over a broad set of use cases. Although it is becoming common to build machine-learned models trained on manual annotations to mitigate such limitations (9), there still remain critical challenges: 1) Scalability: with commercial conversational agents, there are tens of thousands skills available built by 3rd-party developers, and it is prohibitive to collect sufficient amounts of human annotations for all use cases; 2) Discrepancy with actual user satisfaction: annotators do not have full visibility into user's goal and context, and they make their best guess according to the annotation guidelines.

To address these challenges, in this paper, we explore the use of post experience user feedback. As shown in the example dialog below, we instrumented a feedback elicitation system to ask user's feedback with pre-designed prompts such as "Did I answer your question?" and to interpret user's response.

User: When will it stop raining in New York?
Agent: In New York , intermittent rain is possible throughout the day. Did I answer your question?
User: Yes.
Agent: Thanks for your feedback.

Figure 1: An example of an Alexa dialog that asks user feedback.

Unlike the conventional manual annotation, user feedback is a good proxy to user satisfaction as users know best whether a conversational agent provided the right experience they wanted, and the amount of user feedback can easily be scaled up to several orders of magnitude larger than that of human annotation. However, frequent feedback requests can introduce significant friction in user experience, thus we cautiously prompt users with a controlled rate. A machine-learned feedback prediction model is built to produce user satisfaction assessment when direct user feedback is unavailable.

During the course of our early exploration, we identified a few issues in collecting user feedback, which are preventing us from building a holistic metric using the user feedback signal only. Not all scenarios were applicable for collecting user feedback (e.g., when a user barges in and terminats the dialog, asking a post-experience feedback can lead to an unnatural experience.), and not all domains/skills were able to onboard the feedback collection system at the same time. This caused an incomplete coverage of the user feedback data. In contrast, human annotation-based approaches are unobtrusive to user experience and applicable to all situations.

Therefore, in this work, we propose a practical hybrid approach to take the best of both worlds. Our hybrid approach fuses direct user feedback and two types of predicted user satisfaction by two machine-learned models, one trained on user feedback data and the other on human annotation data. During the inference time, a simple waterfall policy is employed for each pair of user utterance and system response: 1) We first check if a direct user feedback is available, which is rare, and respect it if available; 2) Otherwise, we check if the feedback-based prediction model is eligible and its prediction result shows a high confidence score. If it is, we take that feedback prediction; 3) Finally, in case we could not get a prediction from prior stages, we make a prediction with the human annotation-based model. On an Alexa's large-scale test dataset, our hybrid approach achieved significant improvements in precision, recall, F1-score, and PR-AUC (Precision-Recall Area-Under-Curve) by 4.4%, 28.7%, 18.3%, and 24.4%, respectively. Along with this performance improvement, in terms of data volume, the hybrid dataset had less dependency on human annotation as we were able to collect user feedback data at scale. This is another benefit of our hybrid approach.

The rest of the paper is structured as follows. In section 2, we summarize related work. In section 3, we provide a brief analysis to understand the quality and traits of user feedback. In section 4, we present our proposed hybrid approach. In section 5, we describe our experimental setup and results. Finally, in section 6, we conclude with discussions and future work.

2 Related work

One of the conventional approaches to evaluate the quality of intelligent assistant systems is to measure the relevancy of the response of the system using some IR measures such as precision and cumulative gain measures such as NDCG (10; 11). This approach, however, requires human judgement for the relevancy measures, which is generally costly and hard to scale. Such relevancy-based metrics, however, often do not capture the holistic view of system performance such as user satisfaction. To overcome this, some prior works in the search domain have studied users' behavioral patterns to infer their satisfaction level with respect to search results (12; 13; 14; 15). There is also an attempt to understand the relationship between search engine effectiveness and user satisfaction (16). In the area of spoken dialogue system, PARADISE (17) proposed a framework for evaluating goal-oriented dialogues, by specifying the relative contribution of various factors via a linear regression model. For modern intelligent assistants, there was an in-depth study about user satisfaction by classifying the user-system interaction patterns into several categories (18). Another work proposed a research agenda about context-aware user satisfaction estimation for mobile interactions using gesture-based signals (19). These works (20; 21) estimated conversation quality via user satisfaction estimation. However, most prior works are annotation intensive. There is an interesting work that pointed out the issues with annotation-based approaches (22), which aligns with our motivation toward feedback-based user satisfaction estimation approaches, and even further with our hybrid approach. The ability to accurately predict user satisfaction enables a conversational agent to evolve in a self-learning manner. This overview article (3) about personal digital assistants discussed user experience prediction using customer feedback. A recent work on Alexa showed how a conversational agent learns to fix speech recognition and language understanding errors by leveraging an automated user satisfaction predictor without requiring manual supervision (23).

3 User feedback analysis

This section provides an analysis to understand how user feedback correlates with human annotation. The primary dataset used for our correlation study contains 7,447 utterances with user feedback, and the dataset was annotated by trained annotators using the same annotation guidelines of our human annotation process. In the correlation-study dataset, prepared by the annotation work, about 36.5% of utterances are mapped to feedback categories other than YES or NO. An example of the YES feedback category is shown in Figure 1. Among these other categories, the biggest bucket is SILENCE where users did not provide any feedback. Our findings indicate that the majority of SILENCE feedback correspond to satisfying experience. As a more in-depth analysis is required to fully understand other categories, in this work, we decided to utilize only those utterances with a YES or NO feedback which amounts to 4,729 utterances. In our analysis, the user feedback and human annotation had an agreement accuracy of 97.4%, and the Cohen's kappa coefficient (κ) [2] (24) between the user feedback data and the human annotation data was 0.7877 (substantial agreement according to typical kappa interpretation (25)). Note that the high agreement rate between user feedback and human annotation can be partly due to its bias toward satisfaction feedback as we do not have feedback elicitation opportunities when users barge in and ask for termination that are strongly correlated with user's dissatisfaction. To compensate this limitation, our hybrid approach supplements user feedback with human annotation when the use of user feedback is ineligible.

4 Method

This section first describes a deep neural architecture that we designed to build predictive models for both user feedback and human annotation, and then provides the details of our hybrid approach to fuse several inputs of user satisfaction assessment.

[2] *Cohen's kappa coefficient*, `https://en.wikipedia.org/wiki/Cohen%27s_kappa`

4.1 Deep neural model for user satisfaction prediction

Before diving into modeling details, we first introduce a few terminologies. We define a pair of a user utterance and an agent response as a *turn*. Figure 1 shows an example dialog consisting of two turns. The eliciting and user feedback. We call the first turn the *targeted turn* where the conversational agent asks "Did I answer your question?", and the second turn the *answering turn*. Besides the user utterance and agent response, each turn also has meta information such as the timestamp of the turn, the conversational skill invoked to handle the turn, the active screen availability for the turn happened, and etc. [3] Thus, we represent a turn as

$$t^i = (u^i, a^i, f^i) \tag{1}$$

where u^i is user utterance text, a^i is the agent's response text, and $f^i = [f^{i1}, ..., f^{ik}]$ is a list of meta information features, including categorical and numerical features. With the concept of *turn* defined above, in order to capture contextual cues from the surrounding turns such as user's rephrasing patterns and barge-in patterns, we consider another concept of *dialog session* consisting of the targeted, answering, and their surrounding turns. We define a *session* as a maximum span of turns where any two adjacent turns have a time gap less than Δ minutes. In other words, given a long list of turns, we obtain a disjoint set of sessions where time gap between sessions is always greater that Δ minute. Formally, we denote a session as

$$s_j = \{t_j^i\}_{i=1}^{l_j} \tag{2}$$

where l_j represents the number of turns in session j. Our dataset consisting of a set of session and label pairs can then be represented as

$$\{(s_j,\ y_j)\}_{j=1}^{N}$$

where y_j denotes the binary user satisfaction with respect to the targeted turn in session j. Given session information s_j, the model produces a prediction score $p_j \in [0, 1]$ as the probability that the user is *dissatisfied*. Note that we treat dissatisfaction as our primary class since accurately detecting defective experiences offers greater value for us to improve downstream components.

Regarding model input features, we noticed that some meta information features are shared across turns in the same session such as device-type. We treat them as session-level features. With this pre-processing, our model is given five types of features: turn-level textual features, turn-level categorical features, turn-level numerical features, session-level categorical features, and session-level numerical features. To process the session information, the key step is to encode sequential information. Specifically, we need to encode a sequence of words in each turn level. Then, we need to encode a sequence of turns in a session level. In our model, we use a GRU+Attention approach: We scan the sequence with a GRU (26) layer and then use multi-headed attention (27).

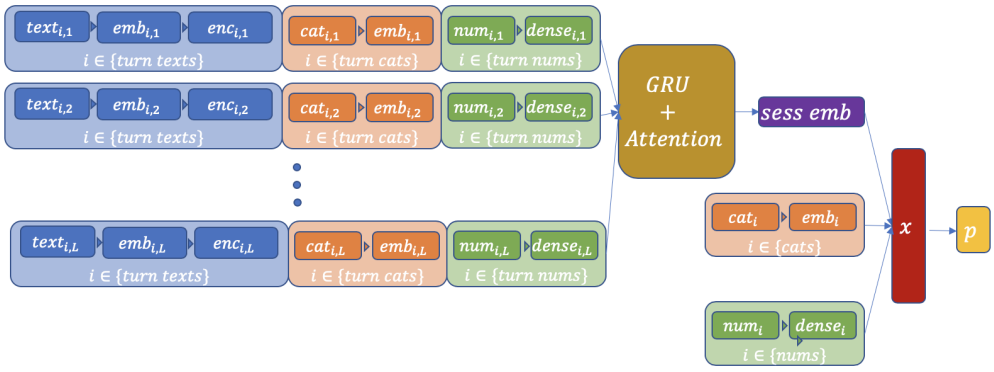

Figure 2: Overall architecture of our deep neural model for user satisfaction prediction: 1) Each turn is represented by turn-level textual features, turn-level categorical features, and turn-level numerical features. 2) Those featurized session turns are fed to the GRU+Attention block and outputs a session embedding. 3) The session level categorical feature and numerical features are encoded, concatenated with the session embedding, and then fed to feed-forward dense layers.

[3] Due to confidentiality reasons, we are not allowed to disclose the exact meta information features.

Taking these features as input, our model employs a deep-wide style neural network (28) to accommodate both structured and unstructured features. Our model has a hierarchical structure to process the word encodings at both turn and session levels. The model architecture is shown in Figure 2.

For each turn, we convert user utterance and agent response texts separately into sequences of word embeddings, using GloVe (29). Each word embedding sequence is fed to a GRU+Attention layer to obtain a sentence embedding. Then a utterance-response embedding is formed by adding these two sentence embeddings. Consequently, a turn representation is a concatenation of the utterance-response embedding and embeddings of turn-level categorical and numerical features. Note that for each numerical feature, we use a single dense layer to obtain a high dimensional encoding while one-hot encodings are used for categorical features.

Given sequence of turn-level representations, we repeat the similar process to produce a session representation by feeding the encoded turns to another GRU+Attention layer, then concatenating the encoded result with session-level categorical feature embeddings and numerical feature embeddings. The final classifier component simply consists of two dense layers. The standard binary cross entropy loss is used as a loss function.

With the deep neural architecture, two user satisfaction models are built: 1) a user feedback prediction model (FP) trained on user feedback data; 2) a fallback prediction model (HP) trained on human annotation data. Note that to make the user feedback prediction model applicable to general cases where feedback turns are non-existent, we remove the feedback prompt and answering turn from the session data for both training and evaluation.

4.2 Hybrid approach

The goal of our hybrid approach is to accurately predict user satisfaction while maximizing the utility of user-agent interaction results (feedback) and its human assessments (annotation) as depicted in Figure 3. Mainly, there are three types of candidate inputs that are captured in the prediction layer, such as explicit user feedback, inferred user satisfaction, and skill-provided assessment:

- **Explicit user feedback** As true user satisfaction is not observable, we ask user for post-experience feedback as the best proxy. This is the most direct method to check whether the experience was satisfying with some caveats: 1) frequent feedback request introduces friction, thus we should cautiously use it with a controlled rate; 2) it is biased toward positive feedback as we do not have an opportunity to collect feedback when there are barge-in and early termination that are strong indicators of negative user experience; 3) its coverage is currently limited to a set of allow-listed experiences.
- **Inferred user satisfaction** User feedback is not always available. A predictive model allows us to measure user satisfaction even when user feedback is not directly collected. To build accurate machine-learned predictors, in the feature layer, we consider various input features such as conversation history, contextual features, domain signals, user profile, historical features, and external knowledge. Specifically, we utilize the FP and HP models previously described in Section 4.1.
- **Skill-provided assessment** While there are several implicit indicators of user dissatisfaction that are skill agnostic, assessing positive user satisfaction often requires knowledge of the target skill and access to skill-specific signals. For example, in the media consumption domain, 30-sec playback is commonly used as an implicit signal of user satisfaction. In a map/navigation application, we may declare success if we see no changes to the destination or route cancellation within a certain time interval (e.g., 15 secs). In a ticket-booking application, we may use a booking confirmation signal, followed by the absence of cancellation within a certain time interval (e.g., 5 mins). Incorporation of skill-provided assessment is out of scope and we leave it for our future work.

As we are at an early stage of leveraging heterogeneous prediction sources, our fusion layer follows a simple waterfall policy to determine whether the conversational agent's action/response was satisfying:

1. If explicit user feedback is present and interpretable, then determine user satisfaction accordingly based on the feedback. Otherwise, go to the next step.

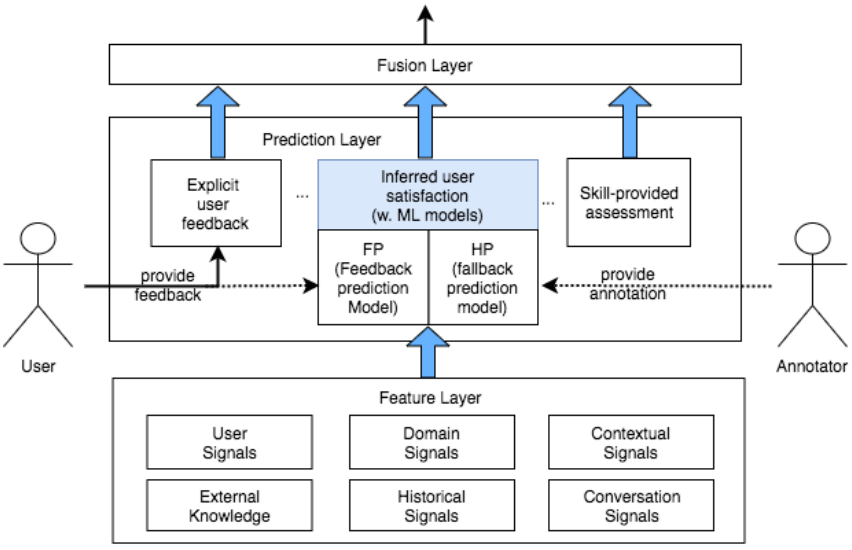

Figure 3: Diagram of the overall architecture of our hybrid model. Both FP and HP models are regularly trained and deployed. The dashed lines (annotation data and feedback data) mean that those data are provided for offline model training. The solid line for feedback signal means its real-time online availability.

2. In case the FP model (the feedback prediction model trained on user feedback data) is eligible for that experience, we run the FP model to infer user satisfaction. If the prediction confidence is high, then determine user satisfaction accordingly based on the FP prediction result. Otherwise, we move to the next step. The confidence threshold is tuned based on a separate development dataset.
3. As a fallback, we predict user satisfaction with the HP model (the fallback prediction model trained on human annotation data). The HP model covers all types of experiences as human annotation does.

5 Experiment

In this section, we describe the datasets and experimental setup, and present the experimental results.

5.1 Datasets

For our experiments, we have allow-listed 43 Alexa intents and collected post experience user feedback at 0.01% sampling rate over Alexa live traffic. We have collected 1.3 million data points which is split into training, validation, and test set with 80%, 10%, and 10% ratios. The training and validation sets were used to train the FP model and the test set was secured to be used only for reporting. For human annotation data, we took a recent chunk of historical Alexa experience annotations that amount to 0.5 million data points. This data set covers all intents and is split into training, validation and test data sets with 80%, 10%, and 10% ratios. Note that the size of feedback dataset is larger than that of human annotation dataset as the feedback collection process is much faster and cheaper than the human annotation process.

Due to the limitations of user feedback coverage, we designed an algorithm to build a composite ground-truth test set which weaves user feedback data with human annotation data. The basic idea is we use collected user feedback as ground truth for traffic segments that are eligible for feedback elicitation and use human annotation for all other traffic segments. The ineligible traffic segments include the following:

- Intents that are not allow-listed for feedback elicitation.
- Turns with barge-in or termination request from users.

- Turns with Unhandled requests, where the conversational agent saying "I'm sorry..." or no response.
- Turns with user feedback values other than YES or NO.

We compute the proportions of each case in the live traffic and bring in human annotation data for the corresponding amounts. The detailed ground-truth test set construction algorithm is listed in Algorithm 1.

Algorithm 1: Building ground truth dataset

input : $\{N_s\}$ //N_s is the number of ground truth examples for segment s
$\{H_s\}$ //H_s is a set of human annotated examples for segment s; H_s^f denotes a subset of examples that could be eligible for user feedback and $H_s^{\neg f}$ ineligible
$\{F_s\}$ //F_s is a set of user feedback examples for segment s
$\{R_s^i\}$ //R_s^i is an ineligible rate for user feedback for segment s
$\{R_s^o\}$ //R_s^o is a rate of receiving user feedback categories other than YES/NO for segment s
S //S is the entire set of segments
S_f //S_f is a set of segments allow-listed for user feedback

output : $\{G_s\}$ //G_s is a set of ground truth labels for segment s

for $s \leftarrow S$ **do**
  if *s not in S_f* **then**
    $G_s \leftarrow$ sample N_s examples from H_s;
  else
    $n_s^{hi} = R_s^i * N_s$;
    $n_s^{ho} = R_s^o * (N_s - n_s^{hi})$;
    $n_s^f = N_s - n_s^{hi} - n_s^{ho}$;
    $G_s^{hi} \leftarrow$ sample n_s^{hi} examples from $H_s^{\neg f}$;
    $G_s^{ho} \leftarrow$ sample n_s^{ho} examples from H_s^f;
    $G_s^f \leftarrow$ sample n_s^f examples from F_s;
    $G_s = G_s^{hi} \cup G_s^{ho} \cup G_s^f$;
  end
end

5.2 Experiment results

Our experiment results are presented in Figure 4. To demonstrate the effectiveness of leveraging user feedback, we compare the following three approaches: 1) *HP*: solely relying on the HP model (annotation-based model), 2) *EFB + HP*: fusion of explicit user feedback and HP model prediction according to our hybrid approach and 3) *EFB + FP + HP*: fusion of explicit user feedback, FP model prediction (feedback-based model) and HP model prediction according to our hybrid approach. Note that whenever users provide explicit user feedback, our hybrid approach takes it as output instead of making any inference (based on the fusion approach described above), meaning whenever a user feedback is explicitly given, our hybrid approach can trivially make the right prediction. Thus, evaluating our hybrid approach requires a parameter that controls the rate at which we assume users provide explicit user feedback. Specifically, given a feedback collection rate, we mark user feedback instances as "given by user" in the ground-truth test data until the rate is met. Then for the marked instances, our hybrid approach takes the associated user feedback as its prediction. In our experiments, we varied feedback collection rate to have a value among the following: 0.01%, 1%, and 10%. The value of 0.01% is the actual feedback collection rate we chose for the live feedback collection for the experiment, and the other two rates are hypothetical for projective purposes.

The *micro-averaged* [4] overall performance results at feedback collection rate of 0.01% clearly demonstrate a large gain that our proposed hybrid approach (i.e. *EFB + FP + HP*) brings in. Compared to a conventional approach (i.e. *HP*, a prediction model trained on human annotation data),

[4] *Micro-aveaged* metrics were calculated with all samples in the test set.

	Feedback Collection Rate (0.01%)							
	Micro Average				Macro Average (Std) across Domains			
Method	Precision	Recall	F1_score	PR_AUC	Precision	Recall	F1_score	PR_AUC
HP	0.7715	0.4656	0.5807	0.6433	0.8155 (0.1254)	0.5444 (0.2607)	0.6225 (0.2261)	0.7029 (0.2218)
EFB+HP	0.7715	0.4657	0.5808	0.6433	0.8156 (0.1253)	0.5445 (0.2606)	0.6226 (0.2260)	0.7030 (0.2217)
EFB+FB+HP	**0.8056**	**0.5991**	**0.6872**	**0.8005**	**0.8401 (0.0675)**	**0.6013 (0.2120)**	**0.6817 (0.1658)**	**0.7788 (0.1421)**

	Feedback Collection Rate (1%)							
	Micro Average				Macro Average (Std) across Domains			
Method	Precision	Recall	F1_score	PR_AUC	Precision	Recall	F1_score	PR_AUC
HP	0.7715	0.4656	0.5807	0.6433	0.8155 (0.1254)	0.5444 (0.2607)	0.6225 (0.2261)	0.7029 (0.2218)
EFB+HP	0.7742	0.2703	0.5851	0.6485	0.8199 (0.1162)	0.5470 (0.2573)	0.6206 (0.2207)	0.7063 (0.2167)
EFB+FB+HP	**0.8071**	**0.6023**	**0.6898**	**0.8032**	**0.8417 (0.0657)**	**0.6030 (0.2103)**	**0.6836 (0.2104)**	**0.7807 (0.1402)**

	Feedback Collection Rate (10%)							
	Micro Average				Macro Average (Std) across Domains			
Method	Precision	Recall	F1_score	PR_AUC	Precision	Recall	F1_score	PR_AUC
HP	0.7715	0.4656	0.5807	0.6433	0.8155 (0.1254)	0.5444 (0.2607)	0.6225 (0.2261)	0.7029 (0.2218)
EFB+HP	0.7951	0.5083	0.6202	0.688	0.8400 (0.0836)	0.5653 (0.2360)	0.6491 (0.1894)	0.7281 (0.1896)
EFB+FB+HP	**0.8211**	**0.6266**	**0.7108**	**0.8248**	**0.8513 (0.0542)**	**0.6149 (0.1992)**	**0.6964 (0.1495)**	**0.7938 (0.1287)**

Figure 4: Evaluation results on the ground-truth test data set with varying feedback collection rates.

precision, recall, F1-score, and PR-AUC (Precision-recall area under curve) metrics are improved by 4.4%, 28.7%, 18.3%, and 24.4%, respectively. Looking at the results of *HP* and *EFB + HP*, it is worth mentioning that explicit user feedback barely moves metrics at our current feedback collection rate as the small amount of collected user feedback is easily diluted by the enormous amount of traffic covered by the HP model. This, in turn, signifies the generalization power that the FP model offers, beyond the sparse user feedback samples, enabling us to make accurate predictions for those experiences that are eligible for feedback elicitation but not triggered for elicitation.

The proposed hybrid model also outperforms the other approaches in *macro-averaged* [5] results, as shown in "Macro Average (Std) across Domains" section in tables in Figure 4. To get the *macro-averaged* metrics there, we first calculate domain-level metrics for each domain in the test set. Then, we calculate the *macro-averaged* metrics and their standard-deviations using the domain-level metrics. To alleviate an excessive skew toward long tails of small-volume domains, we selected top 20 domains that covered $\sim 98\%$ of traffic volume. The smallest standard deviation of the proposed hybrid approach indicates that our approach predicts user satisfaction in a more consistent manner across domains than the other approaches which is a critical property to allocate fair amounts of traffic to each domain according to their service quality.

With two hypothetical feedback collection rates at 1% and 10%, one can clearly see how the increased amount of explicit user feedback impacts the accuracy of our hybrid approach, as shown in the middle and bottom tables. As expected, as we collect more feedback, our hybrid approach makes more accurate predictions and performs in a more consistent fashion across all the domains. Although a blind increase of feedback elicitation rate can cause significant friction in user experience, higher feedback elicitation rates can be safely applied to some targeted segments of traffic without the risk of incurring bad user experience.

6 Conclusion

We proposed an effective hybrid approach that outperforms conventional approaches that are solely based on human annotation in the user satisfaction prediction problem. We started from the limitations of the approaches based on human annotation, which motivated us to utilize direct user feedback. Utilizing user feedback is not only more direct in capturing user satisfaction, but also more scalable and cost-effective. Our hybrid approach fuses explicit user feedback, user satisfaction predictions inferred by two machine-learned models, one trained on user feedback data and the other human annotation data, via a waterfall policy. The hybrid approach resulted in significant improvements in performance metrics, and it also achieved the most consistent performance across the domains, which is another strength. Our proposed approach has been verified with Alexa, and we believe the approach can be extended to other conversational system and text-based chatbot applications. We will extend the fusion layer of our hybrid approach by leveraging machine learning methods.

[5] *Macro-averaged* metrics (with standard-deviations) were calculate using domain-level statistics.

References

[1] M. F. McTear, "Spoken dialogue technology: enabling the conversational user interface," *ACM Computing Surveys*, vol. 34, no. 1, 2002.

[2] Google, "Teens use voice search most, even in bathroom," *google's mobile voice study finds*, 2015. [Online]. Available: http://prn.to/1sfiQRr

[3] R. Sarikaya, "The technology behind personal digital assistants: An overview of the system architecture and key components," *IEEE Signal Processing Magazine*, vol. 34, 2017.

[4] M. I. Hwang and R. G. Thorn, "The effect of user engagement on system success: A meta-analytical integration of research findings," *Information Management*, vol. 35, no. 4, pp. 229 – 236, 1999.

[5] W. H. Delone and E. R. McLean, "Information systems success: The quest for the dependent variable," *Information systems Research*, vol. 3, no. 1, pp. 1–95, 1992.

[6] L. A. Kappleman and E. R. McLean, "The respective roles of user participation and user involvement in the information system implementation success," in *ICIS Conference Proceedings*, 1991, pp. 339–349.

[7] Y.-B. Kim, D. Kim, A. Kumar, and R. Sarikaya, "Efficient large-scale neural domain classification with personalized attention," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 2214–2224.

[8] Y.-B. Kim, D. Kim, J.-K. Kim, and R. Sarikaya, "A scalable neural shortlisting-reranking approach for large-scale domain classification in natural language understanding," in *Proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) 2018*, 2018, pp. 16–24.

[9] Y. Ling, B. Yao, G. Kohli, T.-H. Pham, and C. Guo, "Iq-net: A dnn model for estimating interaction-level dialogue quality with conversational agents," in *Proceedings of KDD Workshop on Conversational Systems Towards Mainstream Adoption*, 2020.

[10] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of ir techniques," *ACM Transactions on Information Systems*, vol. 20, 2002.

[11] T. Saracevic, P. Kantor, A. Y. Chamis, and D. Trivison, "A study of information seeking and retrieving: I. background and methodology. II. users, questions and effectiveness. III. searchers, searches, overlap," *Journal of the American Society for Information Science*, vol. 39, pp. 161–176; 177–196; 197–216, 1988.

[12] M. Ageev, D. Lagun, and E. Agichtein, "Improving search result summaries by using search behavior data," in *Proceedings of the 36th annual international ACM SIGIR conference on Research and development in information retrieval*, 2013, pp. 13–22.

[13] J. Jiang, A. H. Awadallah, X. Shi, and R. W. White, "Understanding and predicting graded search satisfaction," in *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, 2015, pp. 57–66.

[14] A. Hassan and R. W. White, "Personalized models of search satisfaction," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 2013, pp. 2009–2018.

[15] Y. Kim, A. Hassan, R. W. White, and Y.-M. Wan, "Playing by the rules: mining query associations to predict search performance," in *Proceedings of the 6th ACM International Conference on Web Search and Data Mining*, 2013, pp. 133–142.

[16] M. S. Azzah Al-Maskari and P. D. Clough, "The relationship between ir effectiveness measures and user satisfaction," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007, pp. 773–774.

[17] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella, "PARADISE: A framework for evaluating spoken dialogue agents," in *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*. Madrid, Spain: Association for Computational Linguistics, 1997, pp. 271–280.

[18] J. Kiseleva, K. Williams, J. Jiang, A. H. Awadallah, A. C. Crook, I. Zitouni, and T. Anastasakos, "Understanding user satisfaction with intelligent assistants," in *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, 2016, pp. 121–130.

[19] J. Kiseleva and M. de Rijke, "Evaluating personal assistants on mobile devices," in *Proceedings of the 1st International Workshop on Conversational Approaches to Information Retrieval*, 2017.

[20] P. K. Bodigutla, L. Polymenakos, and S. Matsoukas, "Multi-domain conversation quality evaluation via user satisfaction estimation," in *NeurIPS 3rd Conversational AI Workshop*, 2019.

[21] P. K. Bodigutla, L. Wang, K. Ridgeway, J. L. S. Joshi, A. Geramifard, and S. Matsoukas, "Domain-independent turn-level dialogue quality evaluation via user satisfaction estimation," in *Proceedings of SIGDial 2019 Conference*, 2019.

[22] L. Aroyo and C. Welty, "Truth is a lie: Crowd truth and the seven myths of human annotation," *AI Magazine*, vol. 36, no. 1, 2015.

[23] P. Ponnusamy, A. Roshan-Ghias, C. Guo, and R. Sarikaya, "Feedback-based self-learning in large-scale conversational ai agents," in *Proceedings of the 32th Annual Conference on Innovative Applications of Artificial Intelligence (IAAI)*, 2020.

[24] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, pp. 37–46, 1960.

[25] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, pp. 159–174, 1977.

[26] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *arXiv:1406.1078*, 2014.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 5998–6008. [Online]. Available: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[28] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, and M. Ispir, "Wide & deep learning for recommender systems," in *Proceedings of the 1st workshop on deep learning for recommender systems*, 2016, pp. 7–10.

[29] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162