Interpreting Voice Assistant Interaction Quality From Unprompted User Feedback

Pragati Verma
Amazon Alexa
vpragati@amazon.com

Sudeeksha Murari
Amazon Alexa
sud@amazon.com

Abstract

Assessing the quality of a task performed by an Intelligent Voice Assistant (IVA) system such as Alexa, Siri, etc. is vital for maintaining a high bar for Customer Experience (CX) with the system. In this paper, we propose an approach to determine the quality of an IVA utterance using a 'feedback' utterance that is interpretable and scalable. Basing the IVA quality assessments on user feedback in a scalable manner helps the AI systems address problems important to users, bridges the gap between qualitative and quantitative measurements. We propose inexpensive techniques to make quality assessments available to IVA components with low latency, such that the downstream interactions can use the context to avoid negative CX.

1 Introduction

IVA systems are a combination of many individual system components. Simply relying on the outputs of these components to make a binary assessment of the quality of an IVA experience, is challenging. There are models trained to detect defects automatically [1, 2, 3, 4] that collect outputs from IVA components and other session information to assess IVA Interaction Quality (IQ). However, such methods use supervised training involving expensive and non-scalable human-annotation processes. Some users may have concerns about human-handling of data in these methods.

In this paper, we develop a system to collect feedback signals that customers provide voluntarily indicating their satisfaction with an interaction, which we call a Unprompted Feedback Detection (UFD) system. As examples, an angry "just play my music!" indicates that the previously rendered experience was dissatisfactory, or a content "thanks!" indicates that the previously rendered experience was good. We posit that the presence of abnormal sentiment in utterances occurring within 60 seconds (empirically determined) of a previous utterance is directly indicative of the quality of a previous experience, thereby making such utterances 'feedback' utterances. The polarity of such feedback utterances would reflect on the previous utterance IQ. Formulating the problem of detecting IVA IQ as an anomaly detection problem enables the use of unsupervised methods without compromising on result quality. This reduces the need for human annotation, making the solution robust, while protecting the privacy of the customers. Often, we find that IVA interactions cannot be classified into a binary rating of success or no success. In this paper, we use a continuous scale representing the anomaly intensity, which represents the reliability of feedback attribution. Feedback with higher confidence score represents higher confidence in attribution to systemic IVA successes or failures. We also quantize the confidence scores for interpretability.

2 Background

In this paper, we use sentiment scores calculated on the feedback utterance using the method proposed by Kim et al. [5] . These scores are available across three dimensions- activation, valence,

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

and satisfaction. While Kim et al. [5] establish correlation between utterance-level sentiment and conversation-level user satisfaction, the question of reliability of such an attribution remains unexplored. We posit that the more 'different' a user's sentiment is from common IVA interaction sentiment, the more reliable it is for assessing previous IQ. To the best of our knowledge, this is the first work that explores methods to establish confidence in the quality attributed to a previous IVA interaction by interpreting the differentness and polarity of a feedback utterance. While any unsupervised anomaly detection technique can be used for this purpose, we use the Isolation Forests (IF) model [6] for our proof-of-concept, which proposes that anomalies are 'few and different' from normal data points and thus easier to 'isolate'.

3 Methodology

3.1 Data

For this paper, we use an internal dataset comprised of 30 days of de-identified customer utterances for a leading IVA system in production. We use this randomized collection of time-series utterance data to build our dataset D comprising of a set of ordered pairs of utterances $(u_0, u_1)_i$. Here each $(u_0, u_1)_i$ correspond to consecutive utterances for some customer, such that the time delay between the two utterances (u_0, u_1) is at most δ. We create a smaller dataset D_{sample} to use in this paper using a stratified sampling approach, using the week number in the timestamp of the utterance pairs as strata. We use $\delta = 60$ seconds as a result from a separate internal analysis. The satisfaction and valence sentiment are added as attributes for each u_1 in D_{sample} using the method proposed by Kim et al. [5].

The Response Quality (RQ) of an IVA response is a measure of the overall quality of the response rendered by the IVA [7]. We collect the IVA responses for all u_0 in D_{sample}, and obtain human-labeled RQ for all such responses using the annotation scheme proposed by Bodigutla et al. [8]. These RQ ratings are available on a 1-5 scale, and also grouped into binary (satisfactory/ dissatisfactory) labels l such that ratings 1,2 are labeled dissatisfactory, and ratings 3-5 are labeled satisfactory. Thus we create another dataset R containing pairs $(u_0, l)_i$, where each u_{0i} in R is the same as u_{0i} in D_{sample}, and l_i is the binary RQ rating of the IVA response to u_{0i}. We use dataset D_{sample} in our experiments, along with dataset R for evaluation as described in Section 3.3.

3.2 Using anomaly detection to identify feedback

The objective of the UFD system built to detect feedback utterances is to identify sentiment that is anomalous, given the usual range of sentiment customers usually have while interacting with IVA systems. As the activation scores are not correlated with customer satisfaction [5], we use valence and satisfaction sentiment scores as inputs. These scores are real-valued with range from -3 to 3. The UFD runs the IF algorithm on the input and produces three outputs: a) the real-valued confidence in the feedback detection (0-1), b) a quantized bin denoting the confidence level (high, medium, low), and c) the polarity of the feedback (positive, negative). In case an instance is identified as feedback, the UFD converts the raw anomaly scores output from the IF model into feedback confidence scores in the range of 0 to 1 by use of a simple min-max scaling transformation. The UFD also hosts a tunable quantizer that outputs confidence bins (high, medium, low) from the real-valued confidence scores. As a result of a separate internal analysis, we use the polarity of the satisfaction sentiment score as the feedback polarity.

3.3 Evaluation criteria

For evaluating the results of our UFD system, we use the application of the Affective Events Theory (AET) on natural language which states that humans respond negatively when negative things happen to them [9, 10] to establish the IVA RQ as a proxy metric for IQ. We use the dataset R described in Section 3.1 to get the binary RQ l of each u_0 in D_{sample}. Next, we assign output labels l_{out} to each u_0 in D_{sample} using the outputs from our UFD. A u_0 is labeled as satisfactory if u_1 is identified as positive feedback, and dissatisfactory if u_1 is identified as negative feedback. For non-feedback utterances in D_{sample}, we do not assign a label to u_0. As a first step, the UFD system needs to be optimized on precision, so that the outputs can be used directly for different applications. As such, we use precision of l_{out} against the binary RQ l as the classification metric for both positive and negative

feedback class detected by the UFD system. We calculate this precision for the high, medium, and low bin separately, to establish that increasing feedback confidence results in increasing precision against the binary RQ, thus demonstrating the reliability of the continuous feedback scale.

4 Experiments and results

4.1 Baseline

Kim et al. [5] establish that there is correlation between utterance-level satisfaction sentiment and conversation-level customer satisfaction. In the absence of the UFD system, the polarity of the satisfaction score of an utterance can be interpreted as feedback on a previous utterance, given the two utterances occur sufficiently close. We map the satisfaction sentiment polarity of each u_1 in D_{sample} to a binary satisfactory/dissatisfactory rating for u_0. A positive satisfaction sentiment polarity of u_1 is mapped to a satisfactory rating for u_0, and negative satisfaction sentiment polarity of u_1 is mapped to a dissatisfactory rating for u_0. The precision of this simple system against annotated RQ for u_0 is taken as the baseline for our experiments.

4.2 Isolation forests model to detect anomalous sentiment

The IF approach expects the fraction of outliers expected in D_{sample} (contamination factor) as input during training. In a separate study to determine α, we take a small random sample of utterances from the logs of the same IVA as the data mentioned in Section 3.1, but for a different time period. We carry out an exercise to label these as feedback/not feedback. Utterances that contain verbiage indicating the IQ of the previous interaction such as “thanks" or “you are stupid", etc. are labeled as feedback. The fraction of utterances in the sample having a feedback label is taken as the α for use in our models.

In addition to α, the IF approach requires the number of subsamples and the maximum number of features to be drawn from the dataset to train each base estimator. Knowing the α, we set up a weak supervision system to create validation labels for hyperparameter search for our model. We select the satisfaction dimension of the sentiment scores and sort the values in ascending order. Then, we select a fraction $\frac{\alpha}{2}$ of the values each from both ends of values and mark them anomalous. This approach marks a fraction α of the values as anomalous including both positive and negative polarities. This system is a weak supervision system, as it uses only one dimension for labeling and also ignores factors such as instance density. We use area under precision-recall curve as the search criterion, and run ten trials for hyperparameter search using a 10-fold cross validation approach. We select the IF model having the greatest area under precision-recall curve in these trials for our experiments.

The anomaly scores output from the IF algorithm is converted to a feedback confidence score in the range 0-1 for interpretability. We run a min-max scaler on the anomaly scores for each u_1 identified as feedback by the UFD to get the feedback confidence scores. We experiment with different intervals for quantization, and choose intervals that give us the best precision in the high and medium bins while preserving statistical power in each bin. Table 1 presents the confidence levels with the corresponding intervals and the percentage of instances in each bin, along with percentage improvement over baseline precision for the different quantized bins, for satisfactory and dissatisfactory classes. We see that in each case the precision is better than the baseline precision. We also see that for the dissatisfactory class, the precision becomes progressively better with increase in the level of confidence. The same is true for the satisfactory class, except for a slight degradation in the medium confidence bin. The improvement in precision with increase in feedback confidence levels establishes the reliability of the continuous feedback scale. These results corroborate our claim that a) the IQ of a previous interaction can be determined by the presence of anomalous sentiment in the following utterance within 60s, and b) the level of confidence in such an attribution depends on the differentness of sentiment in the feedback utterance from normal IVA utterance sentiment.

5 Use of UFD data for CX assessments

Despite the subjectivity, the user-perceived IQ is meaningful for applications such as personalization. To surface systemic successes or failures, we aggregate the outputs from our UFD into metrics Negative Feedback Rate (NFR) and Positive Feedback Rate (PFR), parametrized by a dimension

Table 1: Details on IF confidence bin intervals

Conf. bin	Conf. range	% of utterances	Improvement over baseline precision (satisfactory)	Improvement over baseline precision (dissatisfactory)
high	[0.63, 1]	18.21%	27.02%	72.34%
medium	[0.31, 0.63)	24.75%	3.96%	43.86%
low	[0.0, 0.31)	57.03%	10.15%	36.41%

Table 2: Pearson's correlation between NFR and engagement

Dimension	Pearson's R	p-value
Curated content	-0.694	0.34×10^{-3}
Sports	-0.817	1.15×10^{-7}
Social experiences	-0.462	1.33×10^{-2}
Utilities	-0.447	1.71×10^{-2}

such as domain, user intent, language, etc. These metrics are defined in Equation 1.

$$NFR(dimension) = 1000 \times \frac{n_{neg}}{n_{tot}} \qquad PFR(dimension) = 1000 \times \frac{n_{pos}}{n_{tot}} \tag{1}$$

Here n_{neg} is the number of negative feedback utterances following the IVA dimension, n_{pos} is the number of positive feedback following the IVA dimension, and n_{tot} is the total number of utterances in the IVA dimension. These metrics are reported per thousand for human readability.

As IQ is an important factor driving customer trust in IVA and thus engagement (utterances per day) [11], we discuss the relation between aggregated metrics from UFD data and engagement. In particular, we explore the relation of NFR with engagement in four domains- sports, social experiences, utilities, and curated content. We use dataset D explained in Section 3.1 with our UFD system to identify high-confidence negative feedback utterances for a 30-day period across these four domains to form the raw NFR time series, aggregated by day. We also collect the total number of utterances for each of the four domains in the same 30-day period, aggregated by day to get the raw engagement time series. We transform these time series using a differencing strategy with lags determined empirically, to make these time series stationary. We observe negative correlation significant at the 5% level between transformed NFR and transformed engagement in these explored domains. Table 2 shows the Pearson's correlation coefficient and corresponding p-values for NFR and engagement in the explored domains. Please see Appendix A for more details. The significant negative values for pearson's correlation in Table 2 validate the use of NFR to detect degradations in engagement, and thus CX. Even in the presence of a possible common response variable, the NFR can be used as an observable metric for fluctuations in broader engagement.

6 Conclusion

In this paper, we use an unsupervised approach to provide a continuous measure of UX by detecting user feedback from their interactions with IVA. We establish that the IQ of an IVA task can be determined by the presence of anomalous sentiment, both positive and negative, in a follow-up, or 'feedback' utterance that occurs within 60 seconds of the original utterance. Since IQ is subjective by user, we define metrics NFR and PFR to surface systemic IVA successes and failures, which can be leveraged to improve CX. We also see that NFR can be used as an observable variable for early detection of degradations in user-IVA engagement in explored IVA dimensions. Since engagement is a topline metric used by most IVAs to measure revenue potential, early indicators like the one proposed in this paper are critical. The UFD does not use any human annotation, and as such the UFD data is available essentially 'for free' to the IVA, reducing operational costs on the IVA components.

References

[1] Alexander Schmitt, Benjamin Schatz, and Wolfgang Minker. Modeling and predicting quality in spoken human-computer interaction. *Proc. SIGDIAL 2011 Conf. 12th Annu. Meet. Spec.*

Interes. Gr. Discourse Dialogue, pages 173–184, 2011.

[2] Stefan Ultes and Wolfgang Minker. Interaction quality estimation in spoken dialogue systems using hybrid-HMMs. *SIGDIAL 2014 - 15th Annu. Meet. Spec. Interes. Gr. Discourse Dialogue, Proc. Conf.*, (June):208–217, 2014. doi: 10.3115/v1/w14-4328.

[3] Stefan Ultes, Robert ElChab, and Wolfgang Minker. Application and Evaluation of a Conditioned Hidden Markov Model for Estimating Interaction Quality of Spoken Dialogue Systems. In Joseph Mariani, Sophie Rosset, Martine Garnier-Rizet, and Laurence Devillers, editors, *Nat. Interact. with Robot. Knowbots Smartphones*, pages 303–312, New York, NY, 2014. Springer New York. ISBN 978-1-4614-8280-2.

[4] Praveen Kumar Bodigutla, Aditya Tiwari, Josep Valls Vargas, Lazaros Polymenakos, and Spyros Matsoukas. Joint turn and dialogue level user satisfaction estimation on multi-domain conversations. *arXiv*, (Table 1), 2020. ISSN 23318422. doi: 10.18653/v1/2020.findings-emnlp.347.

[5] Yelin Kim, Joshua Levy, and Yang Liu. Speech sentiment and customer satisfaction estimation in socialbot conversations. *arXiv*, pages 1–5, 2020. ISSN 23318422.

[6] Fei Tony Liu, Kai Ming Ting, and Zhi Hua Zhou. Isolation forest. *Proc. - IEEE Int. Conf. Data Mining, ICDM*, pages 413–422, 2008. ISSN 15504786. doi: 10.1109/ICDM.2008.17.

[7] Andreas M. Klein, Andreas Hinderks, Martin Schrepp, and Jorg Thomaschewski. Measuring User Experience Quality of Voice Assistants Voice Communication Scales for the UEQ+Framework : Voice Communication Scales for the UEQ+ Framework. *Iber. Conf. Inf. Syst. Technol. Cist.*, 2020-June(July), 2020. ISSN 21660735. doi: 10.23919/CISTI49556.2020.9140966.

[8] Praveen Kumar Bodigutla, Longshaokan Wang, Kate Ridgeway, Joshua Levy, Swanand Joshi, Alborz Geramifard, and Spyros Matsoukas. Domain-independent turn-level dialogue quality evaluation via user satisfaction estimation. *arXiv*, 2019. ISSN 23318422.

[9] Howard M Weiss and Russell Cropanzano. Affective Events Theory: A theoretical discussion of the structure, causes and consequences of affective experiences at work. In *Res. Organ. Behav. An Annu. Ser. Anal. essays Crit. Rev. Vol. 18.*, pages 1–74. Elsevier Science/JAI Press, US, 1996. ISBN 1-55938-938-9 (Hardcover).

[10] Kenneth Joseph, Wei Wei, Matthew Benigni, and Kathleen M. Carley. A social-event based approach to sentiment analysis of identities and behaviors in text. *J. Math. Sociol.*, 40(3):137–166, 2016. ISSN 15455874. doi: 10.1080/0022250X.2016.1159206.

[11] Farzaneh Nasirian, Mohsen Ahmadian, and One Ki Daniel Lee. AI-based voice assistant systems: Evaluating from the interaction and trust perspectives. *AMCIS 2017 - Am. Conf. Inf. Syst. A Tradit. Innov.*, 2017-August(May), 2017.

A Appendix

Figure 1 shows the relationship between scaled transformed NFR and scaled transformed engagement in sports, social experiences, curated content and utilities dimensions.

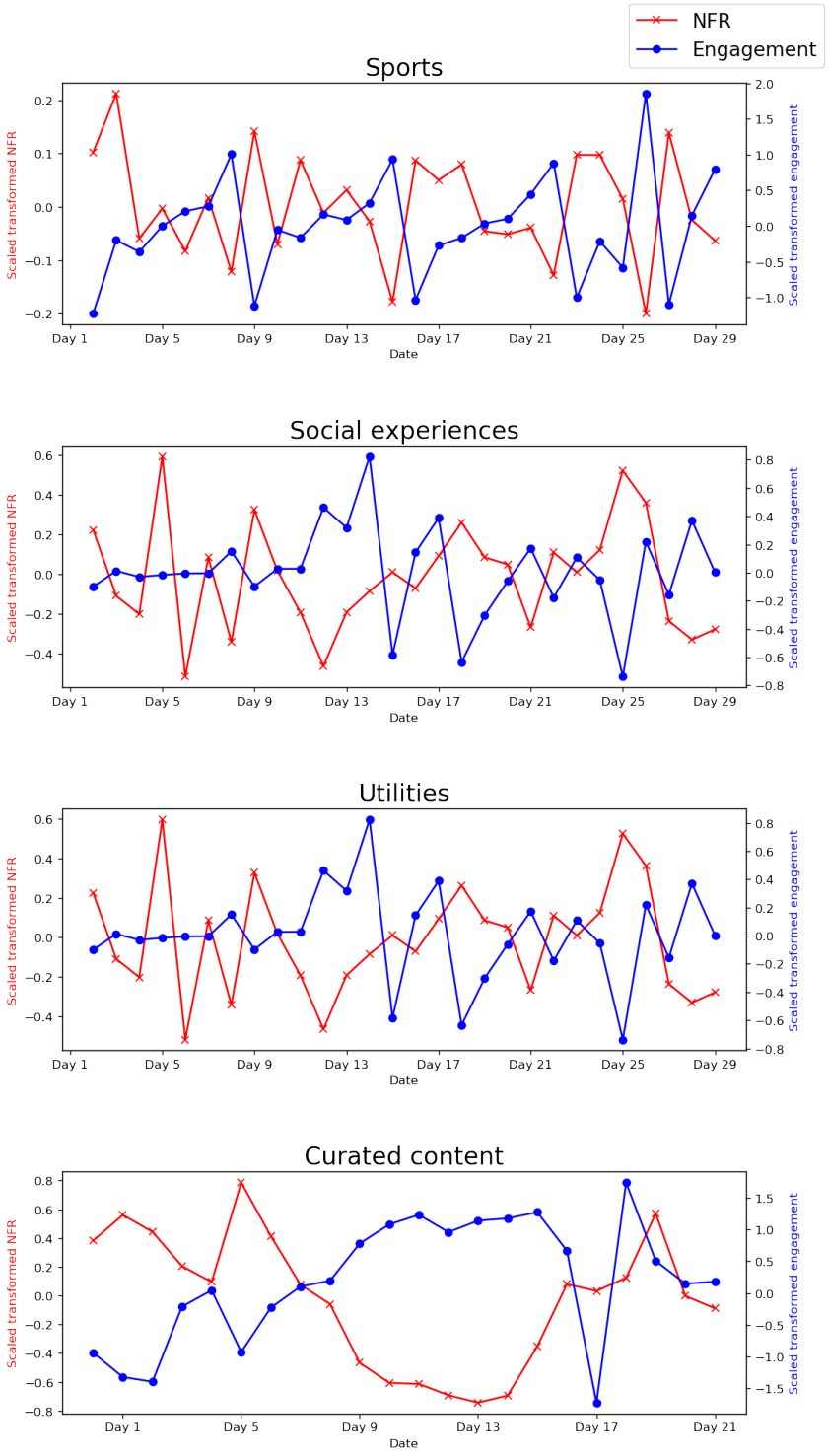

Figure 1: Scaled transformed NFR and Scaled transformed engagement vs date for different dimensions.