Adaptive Multimodal Sentiment Analysis with Stream-Based Active Learning for Spoken Dialogue Systems

Atsuto AJICHI[1], **Takato HAYASHI**[1], **Kazunori KOMATANI**[2], **Shogo OKADA**[1],
[1]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
[2] University of Osaka, Osaka, Japan

{s2410004, hayashi0884}@jaist.ac.jp,
komatani@sanken.osaka-u.ac.jp,
okada-s@jaist.ac.jp

Correspondence: okada-s@jaist.ac.jp

Abstract

In empathic dialogue systems, it is crucial to continuously monitor and adapt to the user's emotional state. To capture user-specific mappings between multimodal behaviors and emotional states, directly asking users about their emotions during dialogue is the most straightforward and effective approach. However, frequent questioning can cause inconvenience to users and diminish the user experience, so the number of queries should be minimized. In this study, we formulate personalized multimodal sentiment analysis (MSA) as a stream-based active learning problem, where user behaviors are observed sequentially, and we assume that the system has an ability to decide at each step whether to request an emotion label from the user. Simulation experiments using a human–agent dialogue corpus demonstrate that the proposed method efficiently improves performance even under few-shot conditions. These results indicate that our approach is effective for developing dialogue systems that achieve cost-efficient personalized MSA.

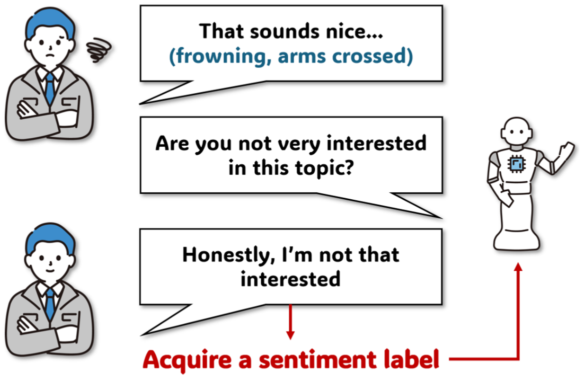

That sounds nice...
(frowning, arms crossed)
Are you not very interested in this topic?
Honestly, I'm not that interested
Acquire a sentiment label

Figure 1: Example of label acquisition process.

1 Introduction

Dialogue systems need the ability to monitor user sentiment and adjust their responses accordingly (Hirano et al., 2019). Sentiments are conveyed not only through verbal cues but also through nonverbal cues such as facial expressions and prosody. By detecting these social signals displaying the sentiment state of the dialogue user, a system can accurately recognize the sentiment state (Vinciarelli et al., 2009), and generate more empathetic responses and provide a richer user experience.

However, sentiment expression patterns vary considerably across individuals. For example, the modality through which sentiments are more prominently expressed and the intensity of such expressions differ from person to person (Binetti et al., 2022; Özer and Göksun, 2020; Kim et al., 2020). Consequently, general models that treat all users uniformly have inherent limitations in estimation performance. Existing studies (Li et al., 2023; Li and Washington, 2024) have demonstrated that personalized models achieve higher accuracy than general ones in sentiment analysis.

For intelligent dialogue systems, directly asking users about their sentiment state is the simplest yet most effective strategy to capture user-specific correspondences between multimodal behaviors and their sentiment as illustrated in Figure 1. Inquiring immediately during dialogue, rather than annotating later from records, reduces the influence of memory decay and yields highly reliable labels. Moreover, this approach eliminates the need for retrospective annotation from recordings or logs, making the process efficient for users. On the other hand, excessive questioning risks degrading the user experience (Komatani and Nakano, 2020). When users are repeatedly asked about their sentiment, they may perceive the interaction as intrusive or unnatural, leading to reduced engagement and willingness to use the system. Thus, it is crucial to design a mechanism that can strategically determine when to query the user for emotion labels, balancing accuracy and user comfort.

Active Learning provides a promising solution to this problem. By selectively querying only the

Proceedings of the 16th International Workshop on Spoken Dialogue Systems Technology, pages 326–337
February 26–March 1, 2026. ©2026 Association for Computational Linguistics

most informative samples, it reduces annotation costs while maintaining model performance (Settles, 2009). In dialogue scenarios, where samples arrive sequentially and labels must be obtained on the spot, stream-based active learning—where the system decides in real time whether to request a label—is more suitable than pool-based active learning. Moreover, recent research on knowledge acquisition through dialogue, such as Waki et al. (2025), has formulated the “when-to-ask” problem using reinforcement learning within a stream-based active learning framework, demonstrating its potential for efficient interactive learning. Thus, in this work, we propose a stream–based active learning framework using reinforcement learning for personalized MSA. Our approach learns policies to decide whether to request labels for sequentially observed multimodal behaviors.

The contributions of this work are threefold:

- We formulate personalized multimodal sentiment analysis as a stream-based active learning problem to address the diversity of individual sentiment expression patterns observed in human-AI interaction.
- We propose a reinforcement learning framework that decides whether to query labels for sequentially observed multimodal behaviors.
- Through experiments on two human–agent dialogue corpora, we demonstrate that the proposed method improves sentiment estimation performance under few-shot conditions.

2 Related Work

2.1 Knowledge Acquisition through Dialogue

One of the essential functions of dialogue systems is the ability to acquire necessary knowledge through interactions with users. Existing models that rely on static knowledge, such as large language models, are often insufficient to cover newly emerging vocabulary, region-specific expressions, or user-specific preferences and affective nuances (Mazumder et al., 2019; Mei et al., 2024). To address this limitation, a growing body of research has explored frameworks that allow knowledge to be incrementally supplemented and updated through user interactions.

Previous studies have investigated knowledge acquisition from various perspectives: acquiring new lexical or factual knowledge from user utterances (Ono et al., 2017; Li et al., 2016), inferring user satisfaction and preferences (Hancock et al., 2019), or enabling robots to learn novel object categories and spatial concepts through dialogue (Taniguchi et al., 2016; Thomason et al., 2019; Kane et al., 2022). Collectively, these works highlight the importance of adapting to the environment and users through actual interactions, rather than relying solely on fixed datasets.

However, frequent questioning in pursuit of knowledge acquisition can negatively affect user experience (Komatani and Nakano, 2020). Therefore, systems must be designed to minimize the number of queries while still efficiently obtaining valuable information. To address this challenge, recent approaches have formulated the acquisition of knowledge and preferences within the framework of active learning, where the system learns "when to ask" (Waki et al., 2025). Such approaches demonstrate the potential for dialogue systems to develop flexible questioning strategies that consider long-term rewards.

2.2 Active Learning for Emotion/Sentiment Recognition

Active Learning (AL) is a framework that improves learning efficiency by selectively requesting labels for the most informative samples from unlabeled data (Settles, 2009). Two representative settings exist: the pool-based setting, where an unlabeled dataset is maintained and samples are selectively queried, and the stream-based setting, where each incoming instance is immediately assessed for its labeling necessity.

AL has also been applied to emotion recognition in order to reduce annotation costs. Li et al. (2024) proposed GRACE, a pool-based AL method that leverages informativeness and cross-modal agreement, demonstrating that high performance can be maintained with limited labeled data. Abdelwahab and Busso (2019) applied AL to speech emotion recognition and showed that performance improvements are achievable with only a small number of labeled samples. More recently, Moreno-Acevedo et al. (2024) introduced a stream-based AL approach that simultaneously considers informativeness and diversity in sample selection, thereby achieving high accuracy with fewer labels. Karnjanapatchara et al. (2024) further integrated multitask learning with annotator agreement modeling, enabling sequential label acquisition while improving reliability.

Nevertheless, most existing studies assume a pool-based setting or rely on a single modality, and relatively little work has focused on methods that can immediately handle sequentially arriving multimodal data. Furthermore, while prior research has emphasized the importance of personalization in adapting to individual differences in emotional expression (Li and Washington, 2024), only limited efforts have explicitly designed AL frameworks with personalization as a primary objective.

Motivated by these gaps, this study proposes a stream-based active learning framework that evaluates whether each sequentially observed multimodal instance is useful for immediate personalization in emotion recognition.

3 Formulation

The goal of this study is to acquire informative feature–label pairs from the early part of a dialogue with a target user and to leverage them for improving sentiment estimation performance in the remaining part. Within this framework, the key challenge is to learn a label query policy that can identify and request only the most useful labels under a limited query budget. Since excessive interruptions may lead to reduced user engagement, it is desirable to achieve substantial performance gains with as few label requests as possible.

We model a dialogue as a sequence of discrete time steps $t = 1, 2, \ldots, T$. At each step, the system observes a multimodal feature $x_t \in \mathbb{R}^d$. The ground-truth label at that step is denoted by $y_t \in \mathcal{Y}$, while the estimated label is denoted by $\hat{y}_t$. The label y_t can be accessed only when the system explicitly queries the user.

The decision to query is governed by a policy π_ϕ, parameterized by ϕ. The reward at each step, r_t, is defined as follows: a positive reward ρ^+ is given when a query prevents a misestimation, a negative reward ρ^- is given when a query is unnecessary, and 0 is given when no query is made. Thus, learning the policy reduces to the following expected reward maximization problem:

$$\max_{\pi_\theta} \mathbb{E}_{\pi_\theta}\left[\sum_{t=1}^{T} r_t\right]. \quad (1)$$

4 Method

In this study, we propose a stream-based active learning method for multimodal sentiment analysis (MSA) to enable real-time personalization. The proposed policy learning framework is based on Reinforced Active Learning (RAL) (Wassermann et al., 2019). Following this framework, we extend it to the multimodal setting by introducing uncertainty quantification based on multimodal features, and propose RAL-MSA, a Reinforced Active Learning approach for MSA. The goal of RAL-MSA is to acquire informative feature–label pairs from the early part of a dialogue with a target user and leverage them to improve sentiment estimation performance in the later part. An overview of the framework is shown in Figure 2.

4.1 Learning Procedure

The overall procedure of RAL-MSA consists of the following four stages:

1. **Pre-training:** Initialize the multimodal classifiers and the label query policy using data from training user data.

2. **Online Adaptation and Label Querying:** Process the target user's data sequentially, one sample at a time. At each step, decide whether to request a label. If a label is requested, compute the reward and update the policy parameters accordingly.

3. **Incremental re-training:** Once the number of newly acquired labeled samples reaches a predefined threshold, add them to the training pool and re-train the classifiers to better reflect the target user's characteristics. Steps (2) and (3) are repeated until the labeling budget is exhausted.

4.2 Overview of Policy Learning

The RAL-MSA framework builds upon RAL (Wassermann et al., 2019). While our overall design follows this paradigm, we extend it to the MSA setting by introducing a new component: *uncertainty quantification based on multimodal features*. This adaptation enables the system to capture uncertainty across heterogeneous modalities (audio, linguistic, visual) rather than relying on unimodal inputs.

The label query policy is therefore learned through three components: (a) uncertainty quantification based on multimodal features, (b) decision-making for label querying, and (c) Policy update via reinforcement learning. These components are designed with the following objectives.

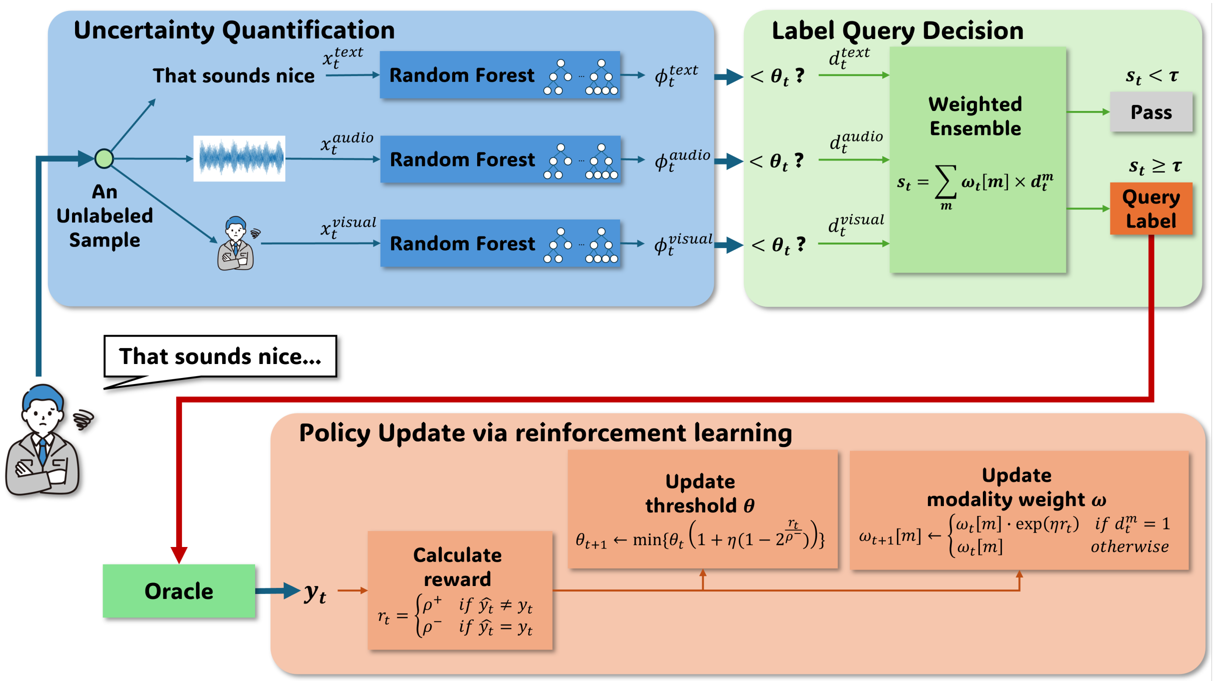

Uncertainty Quantification
That sounds nice
x_t^{text}
Random Forest
ϕ_t^{text}
x_t^{audio}
Random Forest
ϕ_t^{audio}
x_t^{visual}
Random Forest
ϕ_t^{visual}
An Unlabeled Sample
Label Query Decision
$< \theta_t$?
d_t^{text}
$< \theta_t$?
d_t^{audio}
$< \theta_t$?
d_t^{visual}
Weighted Ensemble
$s_t = \sum_m \omega_t[m] \times d_t^m$
$s_t < \tau$
Pass
$s_t \geq \tau$
Query Label
That sounds nice...
Policy Update via reinforcement learning
Update threshold θ
$\theta_{t+1} \leftarrow \min\{\theta_t \left(1 + \eta(1 - 2^{\frac{r_t}{\rho^-}})\right)\}$
Update modality weight ω
$\omega_{t+1}[m] \leftarrow \begin{cases} \omega_t[m] \cdot \exp(\eta r_t) & \text{if } d_t^m = 1 \\ \omega_t[m] & \text{otherwise} \end{cases}$
Oracle
y_t
Calculate reward
$r_t = \begin{cases} \rho^+ & \text{if } \hat{y}_t \neq y_t \\ \rho^- & \text{if } \hat{y}_t = y_t \end{cases}$

Figure 2: Overview of the RAL-MSA method. The oracle is an entity that provides the true label y_t for a queried sample x_t

Uncertainty quantification. The objective is to identify samples with high model uncertainty, for which assigning a ground-truth label is expected to yield the greatest improvement in accuracy. By estimating uncertainty for each modality, the system can assess the confidence of its predictions from multiple perspectives.

Label query decision. Confidence values across modalities are aggregated via a weighted sum to determine whether a label should be requested.

Policy update via reinforcement learning. By updating the query policy in a reinforcement learning framework, the system continuously learns from the outcomes of its own decisions. Based on the rewards, parameters of the query policy are adapted, allowing the system to better account for individual differences in sentiment expression.

4.3 Uncertainty Quantification from Multimodal Features

At each time step t, the input is represented as a multimodal feature vector:

$$x_t = \{x_t^{\text{audio}}, x_t^{\text{text}}, x_t^{\text{visual}}\}. \tag{2}$$

Each modality feature x_t^m is passed through a modality-specific Random Forest model that outputs a predictive distribution over sentiment labels $\mathcal{Y} = \{1, \ldots, C\}$. To estimate uncertainty, the maximum probability is taken as the confidence score for modality m:

$$\phi_t^m = \max_{c \in \mathcal{Y}} P(y = c \mid x_t^m). \tag{3}$$

By evaluating in each modality independently, the system can assess its confidence in samples from diverse viewpoints.

4.4 Label Query Decision

For each modality m, the confidence score ϕ_t^m is compared against a threshold θ_t, and an indicator variable $d_t^m = \mathbb{I}[\, \phi_t^m < \theta_t \,]$ is defined. Only modalities with confidence below θ_t are considered in the decision process, since low-confidence predictions are more likely to correspond to unexplored or ambiguous regions of the feature space.

These indicators are aggregated using modality weights $\omega_t[m]$ to compute a Label request score:

$$s_t = \sum_m \omega_t[m]\; d_t^m. \tag{4}$$

A query is issued if $s_t \geq \tau$. Following RAL (Wassermann et al., 2019), we set the threshold τ to 0.5.

In addition, an ϵ-greedy mechanism queries with probability ϵ whenever a uniform random variable $u \sim \mathcal{U}[0, 1]$ satisfies $u < \epsilon$, even if $s_t < \tau$. This prevents the policy from prematurely ceasing to query and enables detection of misclassified but high-confidence cases as well as novel patterns.

4.5 Policy update via reinforcement learning

The usefulness of each query is reflected in the reward r_t:

$$r_t = \begin{cases} \rho^+, & \text{if a query is made and } \hat{y}_t \neq y_t, \\ \rho^-, & \text{if a query is made and } \hat{y}_t = y_t, \\ 0, & \text{if no query is made.} \end{cases} \quad (5)$$

When the model makes an incorrect prediction and the system requests a label, a positive reward ρ^+ is given. Conversely, when the model's prediction is already correct but the system still requests a label, a negative reward ρ^- is assigned. This design encourages the system to request labels only when doing so is expected to improve performance.

Note that exploratory queries triggered solely by ϵ-greedy sampling are not used for policy updates. Updates are applied only when the committee decision ($s_t \geq \tau$) supports querying.

The confidence threshold is updated following the approach of RAL (Wassermann et al., 2019), as follows:

$$\theta_{t+1} \leftarrow \min\left\{\theta_t\left(1 + \eta\left(1 - 2^{\frac{r_t}{\rho^-}}\right)\right),\ 1\right\}, \quad (6)$$

The θ decreases rapidly when $r_t = \rho^-$, indicating that the system queries too often and needs to be more conservative, while it increases slightly when $r_t = \rho^+$ to acknowledge that the query was beneficial without making the system overly reactive.

The weight of each modality that supported the query ($d_t^m = 1$) is updated multiplicatively:

$$\omega_{t+1}[m] \leftarrow \begin{cases} \omega_t[m] \cdot \exp(\eta\, r_t), & \text{if } d_t^m = 1, \\ \omega_t[m], & \text{otherwise.} \end{cases} \quad (7)$$

The decision power of each modality is reinforced when its judgment is aligned with the full-modality model and the query proves informative; otherwise, its weight is penalized. Finally, the weight vector is normalized to the probability simplex:

$$\omega_{t+1}[m] \leftarrow \frac{\omega_{t+1}[m]}{\sum_{m' \in M} \omega_{t+1}[m']} \quad \forall m. \quad (8)$$

The learning rate η smooths these dynamics so that both θ and ω evolve gradually, avoiding drastic changes from a single query.

4.6 Classification Model

The classification model is trained independently of the label query policy. In this study, we used an ensemble of Random Forest models, as it showed the best performance among the models compared in Experiment A.1. At each time step t, every modality-specific model outputs a predictive distribution $P(y \mid x_t^m)$. The final estimated label $\hat{y}_t$ is then determined by taking the class with the item with the maximum average probability across modalities:

$$\hat{y}_t = \arg\max_{c \in \mathcal{Y}} \frac{1}{|M|} \sum_{m \in M} P(y = c \mid x_t^m). \quad (9)$$

5 Experiments

We compare our method against existing baselines under identical conditions. Below, we describe the datasets, evaluation protocol, and model parameters.

5.1 Datasets

We use two multimodal human–agent dialogue datasets, Hazumi1902 and Hazumi1911 (Komatani et al., 2021; Komatani and Okada, 2021). Hazumi1902 contains dialogues from 28 participants and Hazumi1911 from 26, totaling 4,805 exchanges. Each exchange is annotated with a three-class sentiment label (negative:0, neutral:1, positive:2) derived from self-reported sentiment scores. Further details, including recording conditions are provided in Appendix A.2. Additionally, in this study, we used the participants' text features, audio features, and visual features as input features. Detailed descriptions of each modality are provided in Appendix A.3.

5.2 Evaluation Protocol

We adopt a group-wise cross-validation scheme. All participants are divided into five groups (Groups 1–5). For each experiment, one group is designated as the target group, while the remaining four groups are used for pre-training the classifiers and label query policy. For example, when Group 1 is the target, the models are pre-trained using the data from Groups 2–5.

After pre-training, stream-based active learning is conducted for each participant within the target group using the pre-trained models. For each participant, the first 80% of the dialogue in temporal order is used as the adaptation set, where label queries are issued and the model is updated, while the remaining 20% is held out as the test set. Label requests are issued sequentially (one sample at a time), and the model is retrained whenever five

Table 1: Comparison of Active Learning strategies on Hazumi1902 (n = 28) and Hazumi1911 (n = 25). Values indicate mean Balanced Accuracy (± 95% confidence intervals). Bold indicates the best performance within each condition.

	Hazumi1902			Hazumi1911		
Method	0-shot	5-shot	10-shot	0-shot	5-shot	10-shot
Random Sampling	0.467 ± 0.083	0.468 ± 0.083	0.470 ± 0.080	0.485 ± 0.079	0.477 ± 0.081	0.470 ± 0.086
w/o Threshold (θ fixed)		0.475 ± 0.078	0.471 ± 0.084		**0.491** ± 0.081	0.484 ± 0.078
w/o Weight (ω fixed)		**0.476** ± 0.078	**0.472** ± 0.085		0.490 ± 0.082	**0.485** ± 0.082
Ours (RAL-MSA)		**0.476** ± 0.078	**0.472** ± 0.085		0.490 ± 0.082	**0.485** ± 0.082

newly labeled samples accumulate. The sentiment estimation performance is computed for each participant and averaged across all participants. In this study, we evaluated the models under the 5-shot and 10-shot settings, as in real-world scenarios the system can practically query users only a limited number of times (approximately five to ten) before it becomes intrusive.

We report performance using Balanced Accuracy (BA), which is robust to class imbalance in multiclass classification.

5.3 Hyperparameters

All hyperparameters used in the classifiers and the label query policy were tuned in preliminary experiments and fixed across all runs. Detailed parameter values (e.g., the number of trees, learning rates, and reward settings) are provided in Appendix A.4.

5.4 Comparison Models

To validate the effectiveness of the proposed RAL-MSA, we compare it with the following models.

Random sampling. As a naive baseline, labels are queried by selecting samples uniformly at random from the dialogue stream. This method provides a lower bound for performance, clarifying the benefit of active learning over chance.

w/o threshold θ**.** A variant where the confidence threshold is not adapted online but fixed to the value determined from pre-training data. Since the pre-training data consists of dialogues from multiple non-target users, this value can be regarded as the parameter optimized for an average user. This comparison highlights the importance of individual adaptation of threshold θ.

w/o weight ω**.** A variant where modality weights are fixed to the values obtained from pre-training data and are not updated during interaction. As with the threshold, these values are estimated from multiple non-target users and thus represent the parameters optimized for an average user. This comparison highlights the importance of individual adaptation of modality weights ω.

Including these models clarifies whether the performance gains of RAL-MSA arise from its personalization mechanisms or simply from relying on parameters tuned for an average user.

6 Results and Discussions

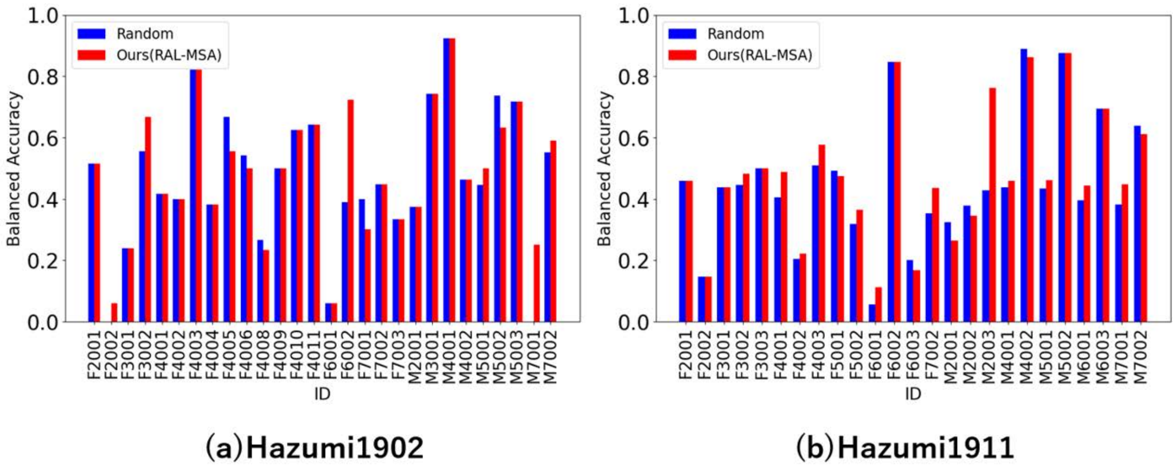
6.1 Comparison of Active Learning Strategy

Table 1 summarizes the performance under the 0-shot, 5-shot, and 10-shot settings for both datasets. One subject (M6002 in Hazumi1911), for whom no samples were queried in the 5-shot setting, was excluded from analysis.

RAL-MSA showed higher accuracy than Random Sampling in both datasets under the 5-shot and 10-shot conditions, by up to 1.5 %. However, in Hazumi1911 (5-shot), the model without threshold adaptation (w/o Threshold) achieved slightly higher accuracy than RAL-MSA. Since RAL-MSA achieved the best results at 10-shot, this suggests that threshold adaptation was not yet effective in the very early learning stage, where limited data may have hindered precise uncertainty calibration. In addition, RAL-MSA and w/o Weight produced the same results across both datasets, implying that the modality-weight adaptation played a minor role in short-term few-shot settings.

In Hazumi1902, the performance of RAL-MSA decreased from 5-shot to 10-shot, whereas in Hazumi1911 both RAL-MSA and the baseline showed a similar decline. Although the cause remains unclear, it may relate to short-term instability or sample selection effects. Future work should test this tendency using larger and longer-term dialogue datasets and adopt label acquisition strategies that account for label diversity to improve stability.

Overall, considerable inter-subject variance was observed, reflecting the validation protocol's sensitivity to individual differences. Further analysis is provided in Section 6.2.



(a)Hazumi1902

(b)Hazumi1911

Figure 3: Performance by ID in 5-shot. The horizontal axis represents participant IDs, and the vertical axis indicates balanced accuracy values. The prefixes “F” and “M” in the IDs denote female and male participants, respectively, followed by two digits indicating the age group. The last two digits serve as identifiers to distinguish participants of the same gender and age group (e.g., F2001 represents a female participant in her twenties).

6.2 Subject-wise Performance Analysis

Figure 3 compares balanced accuracy between Random Sampling and RAL-MSA for each subject in Hazumi1902 and Hazumi1911 (5-shot).

In Hazumi1902, among the 28 subjects, RAL-MSA showed higher accuracy than Random Sampling in 6, underperformed in 5, and achieved identical accuracy in 17. Although most subjects showed comparable accuracy between the two methods, certain individuals (e.g., F6002 and M7001) exhibited more than 20% performance improvement with RAL-MSA. In particular, subjects such as F2002 and M7001, whose Random Sampling accuracy was 0%, achieved non-zero accuracy with RAL-MSA. These findings highlight the effectiveness of active learning in addressing challenges from limited data and individual differences in multimodal behaviors. In Hazumi1911, RAL-MSA showed higher accuracy than Random Sampling in 12 of 25 subjects, underperformed in 6, and tied in 7. Thus, about half of participants benefited from RAL-MSA. In particular, M2003 showed a striking improvement of about 30%, reinforcing the effectiveness of RAL-MSA in adapting to individual uncertainty distributions.

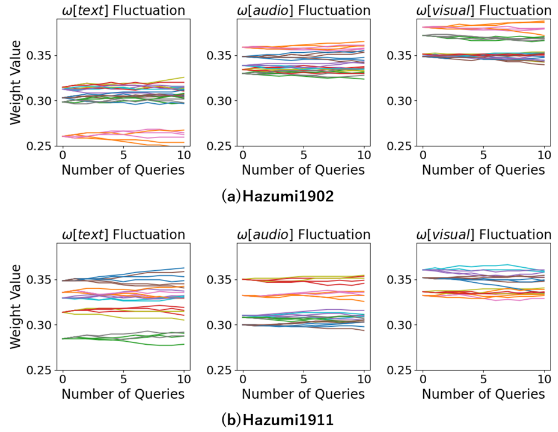
6.3 Analysis of Modality Weight Adaptation

In this section, we analyze modality weight fluctuation to identify the effect of modality weight adaptation. The fluctuation of modality weights during active learning is shown in Figure 4. In this study, the subjects in each dataset were divided into five groups for pre-training (Section 5.2). As a result, five different 0-shot values exist for each dataset. The initial settings for these values are detailed in Appendix A.5. In Hazumi1902, the weight assigned to the visual modality was consistently higher, suggesting that visual cues played a central role in uncertainty estimation for sentiment analysis. A similar trend was observed in Hazumi1911, where the visual weight remained slightly dominant. Interestingly, as the number of queries increased in Hazumi1911, the variance of the text modality weights expanded. This indicates that the importance of linguistic information varied across individuals, implying that the modality weighting mechanism can contribute to personalization by capturing subject-specific differences in uncertainty sources.

In RAL-MSA, if the weight of any single modality exceeds 0.5, that modality alone can satisfy the label-request condition $s_t \geq \tau$. However, as shown in Figure 4, no modality weight exceeded 0.5 during training, suggesting that the weighting mechanism did not directly impact performance. Nonetheless, in longer-term or higher-budget scenarios, adaptive weighting may become more influential as the system accumulates diverse samples and refines modality-specific confidence.

6.4 Analysis of Uncertainty Threshold Adaptation

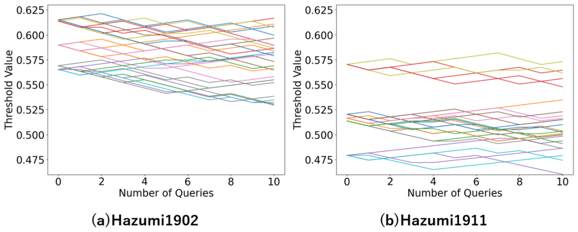
In this section, we analyze uncertainty threshold fluctuation to identify the effect of threshold adaptation. The fluctuation of uncertainty thresholds



(a)Hazumi1902

(b)Hazumi1911

Figure 4: modality weight(ω) fluctuation. The horizontal axis represents the number of queried samples, and the vertical axis indicates the modality weight for each modality. Colors represent each participant.



(a)Hazumi1902

(b)Hazumi1911

Figure 5: threshold(θ) fluctuation. The horizontal axis represents the number of queried samples, and the vertical axis indicates the values of the uncertainty threshold. Colors represent each participant.

over queries is presented in Figure 5. In this study, the subjects in each dataset were divided into five groups for pre-training (Section 5.2). As a result, five different 0-shot values exist for each dataset. The initial settings for these values are explicitly described in Appendix A.5. Both datasets exhibited an increasing intra-fold variance of thresholds over time, indicating that the threshold adaptation mechanism dynamically adjusted according to the uncertainty landscape of each subject.

In combination with the results of the w/o Threshold baseline, this suggests that threshold adaptation was beneficial for personal adaptation, especially beyond the very early stages of training.

6.5 Limitations

As with many dialogue-based personalization studies, this work has several limitations in generalizability. First, the RAL-MSA was evaluated only on two corpora (Hazumi1902 and Hazumi1911) collected under similar conditions, which limits its generalization to diverse users and conversational settings. Future work should examine adaptability to more long-term temporal variations, broader user populations, and corpora in other languages using more diverse datasets.

Second, experiments were conducted in simulation without real users. Hence, practical aspects such as real-time estimation and user experience remain unexplored. Investigating how often queries can be issued without burdening users and how to ensure reliable self-reported labels will be essential for real-world deployment.

Third, the framework assumes queries after each utterance. Exploring more natural timings (e.g., topic boundaries or pauses) and promoting sample diversity are promising directions.

Despite these limitations, this study serves as a first step toward cost-efficient and accurate personalized multimodal sentiment analysis through stream-based active learning.

7 Conclusion

In this study, we proposed RAL-MSA, a reinforced active learning for multimodal sentiment analysis. The RAL-MSA optimizes when to request sentiment labels during ongoing interactions, dynamically adjusting modality weights and uncertainty thresholds to adapt to individual expression patterns while reducing unnecessary queries. Experiments on the Hazumi1902 and Hazumi1911 datasets showed that RAL-MSA performed better than Random Sampling under few-shot conditions, with some participants showing noticeable performance gains. The threshold adaptation appeared to contribute to personalization in later learning stages, and the modality-weight mechanism reflected user-specific differences in information sources. These findings highlight the potential of reinforcement-based active querying for cost-efficient and accurate personalized sentiment analysis. As the evaluation was conducted in a simulation-based setting, the results represent an initial step. Therefore, future work will validate RAL-MSA through real-user experiments in long-term and real-time dialogues to confirm its generality and practical applicability.

8 Acknowledgments

This work was partially supported by JSPS KAKENHI (22H00536, 23H03506), and JST CREST (JPMJCR2563), and JST CRONOS (JPMJCS24K7).

References

Mohammed Abdelwahab and Carlos Busso. 2019. Active learning for speech emotion recognition using deep neural network. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7.

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10.

Nicola Binetti, Nadejda Roubtsova, Christina Carlisi, Darren Cosker, Essi Viding, and Isabelle Mareschal. 2022. Genetic algorithms reveal profound individual differences in emotion recognition. *Proceedings of the National Academy of Sciences*, 119(45):e2201380119.

Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from dialogue after deployment: Feed yourself, chatbot! In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3667–3684, Florence, Italy. Association for Computational Linguistics.

Yuki Hirano, Shogo Okada, Haruto Nishimoto, and Kazunori Komatani. 2019. Multitask prediction of exchange-level annotations for multimodal dialogue systems. In *International Conference on Multimodal Interaction, ICMI 2019, Suzhou, China, October 14-18, 2019*, pages 85–94. ACM.

Benjamin Kane, Felix Gervits, Matthias Scheutz, and Matthew Marge. 2022. A system for robot concept learning through situated dialogue. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 659–662, Edinburgh, UK. Association for Computational Linguistics.

Thus Karnjanapatchara, Sixia Li, Candy Olivia Mawalim, Kazunori Komatani, and Shogo Okada. 2024. Incremental multimodal sentiment analysis for hais based on multitask active learning with interannotator agreement. In *2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 72–79.

Shun Katada, Shogo Okada, and Kazunori Komatani. 2023. Effects of physiological signals in different types of multimodal sentiment estimation. *IEEE Transactions on Affective Computing*, 14(3):2443–2457.

Jangwon Kim, Asterios Toutios, Sungbok Lee, and Shrikanth S. Narayanan. 2020. Vocal tract shaping of emotional speech. *Computer Speech & Language*, 64:101100.

Kazunori Komatani and Mikio Nakano. 2020. User impressions of questions to acquire lexical knowledge. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGdial 2020, 1st virtual meeting, July 1-3, 2020*, pages 147–156. Association for Computational Linguistics.

Kazunori Komatani and Shogo Okada. 2021. Multimodal human-agent dialogue corpus with annotations at utterance and dialogue levels. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8.

Kazunori Komatani, Shogo Okada, Haruto Nishimoto, Masahiro Araki, and Mikio Nakano. 2021. *Multimodal Dialogue Data Collection and Analysis of Annotation Disagreement*, pages 201–213. Springer Singapore, Singapore.

Taku KUDO. 2004. Applying conditional random fields to japanese morphological analysis. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, pages 230–237.

Jialin Li, Alia Waleed, and Hanan Salam. 2023. A survey on personalized affective computing in human-machine interaction. *Preprint*, arXiv:2304.00377.

Jiwei Li, Alexander H. Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. 2016. Learning through dialogue interactions by asking questions. *Preprint*, arXiv:1612.04936.

Joe Li and Peter Washington. 2024. A comparison of personalized and generalized approaches to emotion recognition using consumer wearable devices: Machine learning study. *JMIR AI*, 3.

Xinyu Li, Wenqing Ye, Yueyi Zhang, and Xiaoyan Sun. 2024. Grace: Gradient-based active learning with curriculum enhancement for multimodal sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 5702–5711, New York, NY, USA. Association for Computing Machinery.

Sahisnu Mazumder, Bing Liu, Shuai Wang, and Nianzu Ma. 2019. Lifelong and interactive learning of factual knowledge in dialogues. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 21–31, Stockholm, Sweden. Association for Computational Linguistics.

Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, and Xueqi Cheng. 2024. SLANG: New concept comprehension of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12558–12575, Miami, Florida, USA. Association for Computational Linguistics.

Santiago A. Moreno-Acevedo, Juan Camilo Vasquez-Correa, Juan M. Martín-Doñas, and Aitor Álvarez. 2024. Stream-based active learning for speech emotion recognition via hybrid data selection and continuous learning. In *Text, Speech, and Dialogue*, pages 105–117, Cham. Springer Nature Switzerland.

Kohei Ono, Ryu Takeda, Eric Nichols, Mikio Nakano, and Kazunori Komatani. 2017. Lexical acquisition through implicit confirmations over multiple dialogues. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 50–59, Saarbrücken, Germany. Association for Computational Linguistics.

Demet Özer and Tilbe Göksun. 2020. Gesture use and processing: A review on individual differences in cognitive resources. *Frontiers in Psychology*, 11.

Björn W. Schuller, Stefan Steidl, and Anton Batliner. 2009. The INTERSPEECH 2009 emotion challenge. In *10th Annual Conference of the International Speech Communication Association, INTERSPEECH 2009, Brighton, United Kingdom, September 6-10, 2009*, pages 312–315. ISCA.

Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Akira Taniguchi, Tadahiro Taniguchi, and Tetsunari Inamura. 2016. Spatial concept acquisition for a mobile robot that integrates self-localization and unsupervised word discovery from spoken sentences. *IEEE Transactions on Cognitive and Developmental Systems*, 8(4):285–297.

Jesse Thomason, Aishwarya Padmakumar, Jivko Sinapov, Nick Walker, Yuqian Jiang, Harel Yedidsion, Justin Hart, Peter Stone, and Raymond J. Mooney. 2019. Improving grounded natural language understanding through human-robot dialog. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6934–6941.

Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759. Visual and multimodal analysis of human spontaneous behaviour:.

Issei Waki, Ryu Takeda, and Kazunori Komatani. 2025. Learning to ask efficiently in dialogue: Reinforcement learning extensions for stream-based active learning. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 431–440, Avignon, France. Association for Computational Linguistics.

Sarah Wassermann, Thibaut Cuvelier, and Pedro Casas. 2019. RAL - Improving Stream-Based Active Learning by Reinforcement Learning. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD) Workshop on Interactive Adaptive Learning (IAL)*, Würzburg, Germany.

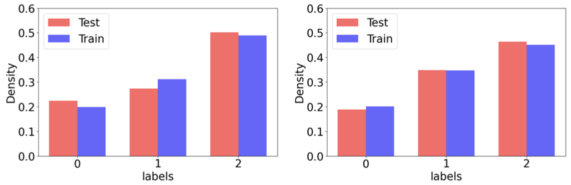
A Appendix

A.1 Model Selection for Sentiment Estimation

Table 2 shows the balanced accuracy of various sentiment estimation models on the Hazumi1902 and Hazumi1911 datasets. Across both datasets, Random Forest achieved the highest performance compared to k-NN, Decision Tree, and MLP classifiers. Based on these results, Random Forest was selected as the sentiment estimation model in this study.

Table 2: Balanced Accuracy (mean [95% CI]) of classification models for multimodal sentiment estimation.Bold indicates the best performance within each model.

Models	**Hazumi1902**	**Hazumi1911**
k-nn	0.424 ± 0.084	0.457 ± 0.075
Decision Tree	0.403 ± 0.063	0.433 ± 0.072
Random Forest	**0.467** ± 0.083	**0.475** ± 0.078
Neural Network	0.425 ± 0.066	0.418 ± 0.075



(a)Hazumi1902

(b)Hazumi1911

Figure 6: Label distribution: The labels 0, 1, and 2 correspond to negative, neutral, and positive classes, respectively.

A.2 Datasets

The Hazumi series consists of corpora of dialogues between human participants and an agent, which is publicly available. Table 3 shows the statistical information of the dataset. For each exchange, participants self-reported their subjective sentiment (SS) using a 7-point scale, where 1 indicates "not enjoying the conversation" and 7 indicates "fully enjoying the conversation." Here, the "exchange" consists of a system utterance followed by a user utterance. Following prior work using Hazumi (Karnjanapatchara et al., 2024), SS scores were converted into three classes: positive (5–7), neutral (4), and negative (1–3). The resulting label distribution is shown in Figure 6.

Dialogues were recorded on video, with the agent controlled by a human operator using the Wizard-of-Oz method. Hazumi1902 contains dialogues from 28 participants (19 female), and Hazumi1911 from 26 participants (14 female).

A.3 Feature Extraction

We extract features from three modalities——audio, linguistic, and visual——strictly following prior

Table 3: Statistics of Hazumi datasets. "exchange" is defined as a unit of data spanning from a participant's utterance to the completion of the system's response.

	Hazumi1911	Hazumi1902
Participants	26	28
Avg. age	44.6 ±16.7	44.6 ±15.2
Avg. duration	20.5 min	17.7 min
Avg. exchanges	95	83
Total duration	534.0 min	495.3 min
Total exchanges	2468	2337

works using Hazumi dataset (Katada et al., 2023).

Text features: Speech transcripts are processed using the Japanese morphological analyzer MeCab (KUDO, 2004). We extract part-of-speech token counts and bag-of-words (BoW) features. Due to differences in vocabulary size, the text feature dimensionality varies across datasets, resulting in 984-dim. for Hazumi1902 and 2613-dim. for Hazumi1911.

Audio features: Using the openSMILE toolkit[1], we extract the INTERSPEECH 2009 Emotion Challenge feature set (IS09) (Schuller et al., 2009) for each utterance, yielding 384-dim. acoustic features such as pitch and energy.

Visual features: Using OpenFace (Baltrušaitis et al., 2016), we track ten facial landmarks and compute frame-wise velocity and acceleration (30 fps) for 12 points around the eyes and mouth. For each turn exchange, we extract the maximum, mean, and standard deviation of these signals, along with averaged AU (action unit) activations. In addition, Microsoft Kinect V2 captures head and shoulder joint motion, from which we compute velocity and acceleration statistics per exchange. The resulting visual feature vector is 86-dim. combining two descriptors: facial expression (66-dim.) and body motion (20-dim.).

A.4 Hyperparameters

For both the classifiers and the label query policy, we use a Random Forest model per modality. Each forest consists of 100 decision trees (n_estimators=100) with no restriction on the maximum depth (max_depth=None), using Gini impurity as the splitting criterion (criterion='gini') and with bootstrapped sampling enabled.

The ϵ-greedy exploration rate in the query policy was set to 0.05. Rewards were assigned as $\rho^+ = 1$ for beneficial queries and $\rho^- = -1$ for redundant ones. The update rate η for both the threshold and modality weights was set to 5×10^{-3} during pre-training and 1×10^{-2} during personalization. The query budget in AL was set to 0.5.

A.5 0-shot values for each subject group

In this study, the subjects in each dataset were divided into five groups for pre-training (Section 5.2). As a result, five different 0-shot values exist for each dataset. The values for each group are shown in Table 4. Values are rounded to the fourth decimal place.

[1] https://www.audeering.com/research/opensmile/

Table 4: 0-shot values for each subject group.

Group	Hazumi1902		Hazumi1911	
	threshold θ	modality weight $\{\omega[\text{text}], \omega[\text{audio}], \omega[\text{visual}]\}$	threshold θ	modality weight $\{\omega[\text{text}], \omega[\text{audio}], \omega[\text{visual}]\}$
Group 1	0.615	{0.303, 0.349, 0.349}	0.520	{0.348, 0.300, 0.352}
Group 2	0.590	{0.260, 0.359, 0.381}	0.516	{0.336, 0.332, 0.332}
Group 3	0.549	{0.300, 0.331, 0.369}	0.514	{0.284, 0.308, 0.408}
Group 4	0.614	{0.315, 0.334, 0.351}	0.570	{0.314, 0.350, 0.336}
Group 5	0.565	{0.313, 0.339, 0.349}	0.479	{0.329, 0.310, 0.360}