Detection of Response Quality Degradation During Crowdsourced Annotation Tasks Using a Conversational Interface

Shigeyuki Taira*, Yuki Matsuda†‡, Yoshinobu Fukumitsu*, Hirohiko Suwa*‡, Keiichi Yasumoto*‡
* *Nara Institute of Science and Technology*, Nara, Japan,
Email: {taira.shigeyuki.tn2, fukumitsu.yoshinobu.ft5, h-suwa, yasumoto}@is.naist.jp.
† *Okayama University*, Okayama, Japan,
Email: yukimat@okayama-u.ac.jp.
‡ *RIKEN Center for Advanced Intelligence Project AIP*, Tokyo, Japan.

***Abstract*—Crowdsourcing-based annotation tasks have been widely utilized as a cost-effective and scalable approach for collecting training data in machine learning. However, the quality of responses obtained through crowdsourcing often varies, posing significant challenges for quality control. To address this issue, this study aims to develop a method for preventing quality degradation by detecting signs of declining response quality in real time. We propose a binary classification model that estimates quality degradation by extracting features from device operation data-such as device orientation and screen interactions-collected during the annotation process and applying supervised machine learning. To evaluate the effectiveness of the proposed method, we developed a smartphone application that supports annotation tasks and continuously collects device operation data in the background. A user study was conducted in which participants were asked to evaluate the correctness of image captions, while their device operation data were continuously recorded during task execution. Using the collected data, we built and evaluated a binary classification model to estimate response quality. The model, trained with time-series validation on intra-individual data, achieved a precision of 0.723, a recall of 0.741, and an F1-score of 0.731. These results suggest that the proposed method is effective in estimating response quality degradation in crowdsourced annotation tasks.**

***Index Terms*—crowdsourcing, annotation, satisficing, careless responses, machine learning**

I. Introduction

Crowdsourcing is a business model in which tasks are delegated to a large, unspecified group of people via the Internet. Due to its advantages in terms of low cost and scalability, it has been widely adopted in various domains. In particular, in the field of machine learning, large-scale training datasets are essential for model development and performance improvement. Crowdsourcing enables the external commissioning of annotation tasks, making it a practical and cost-effective method for acquiring such data. However, a significant challenge in crowdsourcing lies in the variability of data quality. The responses obtained from crowdsourced workers (annotators) are not always reliable, making quality control difficult [1]. This is especially problematic when financial incentives are provided, as annotators may prioritize task completion speed over accuracy in order to maximize their rewards. Moreover, inaccurate responses may also occur unintentionally when the task instructions are not well understood or when annotators are fatigued and lose focus.

Simon *et al.* [2] defined *satisficing* as the tendency of participants to minimize cognitive effort under limited resources, which can degrade machine learning model accuracy if such low-quality responses are included in training data. Real-time detection and prevention are thus essential. In social psychology, attention-check items have been used to detect satisficing [3], [4], but these can stress respondents, reduce motivation, and increase survey burden by adding extra questions. To address these limitations, recent work has explored detecting low-quality responses without adding extra questions. Some research [5], [6] extracted features from smartphone or PC interaction logs after task completion and used machine learning to identify satisficing responses, improving data quality by removing them. However, their method is post hoc, limiting real-time interventions and posing challenges when sample size is small or data representativeness is critical.

In this study, we propose a method to estimate response quality degradation in real time during smartphone-based annotation tasks, specifically focusing on binary image caption verification. The proposed method collects real-time device operation data—including screen interactions such as taps and scrolls, as well as sensor data such as device orientation and acceleration—while the annotation task is being performed. These data are used to extract features and build a binary classification model to estimate quality degradation using supervised learning. To enable this, we developed a smartphone application that supports both annotation task delivery and continuous background logging of device operation data during task execution. We conducted a user study involving university students to collect training data. Using the collected data, we extracted features and constructed binary classification models to estimate the tendency of response quality degradation.

This study was supported in part by JSPS KAKENHI Grant Number JP24K20763.

The model was evaluated from two perspectives: individual-level prediction accuracy and generalization performance. Precision, Recall, and F1-score were used as metrics. For individual-level evaluation, we applied expanding time-series validation, using each participant's earlier tasks as training data and subsequent tasks as test data. For generalization, we excluded one participant's data as the test set and used the rest to build an ensemble model. We collected device operation and response data from 30 participants. Using time-series validation, we built individual models for estimating response quality. The models achieved a Precision of 0.722, Recall of 0.741, and an average F1-score of 0.731. These results suggest the proposed method is promising for real-world use in estimating response quality degradation from smartphone interaction data.

II. Related Work

A. Studies on Detecting Low-Quality Responses

Several studies have been proposed to detect low-quality responses. Well-known examples include the Instructional Manipulation Check (IMC) by Oppenheimer *et al.* [3], and the Attentive Responding Scale (ARS) and Directed Question Scale (DQS) by Maniaci *et al.* [4]. These methods aim to detect satisficing by embedding specific questions in the survey that reveal inattentiveness or inconsistency in the respondent's behavior. While these approaches enable the detection of inattentiveness or dishonesty, these methods may cause psychological stress for respondents by making them feel distrusted. Such items can undermine intrinsic motivation and potentially lead to more low-quality responses. In addition, adding validation items increases the total number of questions, raising the cognitive load on respondents.

Several approaches using sensing technology to identify careless responses have been proposed. Gogami *et al.* [5] introduced a method that eliminates the need for explicit validation items. Instead, they extracted features from smartphone operation logs—such as taps and scrolls—recorded after task completion, and used them to train machine learning models to classify satisficing behavior. However, as a post-hoc method, it only supports reactive measures like discarding unreliable data, making real-time or preemptive interventions infeasible. Fukumitsu *et al.* [6] proposed a real-time detection method for low-quality responses using machine learning. Their approach targets named entity annotation tasks and extracts features from background-logged screen interactions, such as mouse movements and clicks on a PC. These features are used to detect signs of low-quality responses during task execution. Although this method enables real-time detection, it assumes such responses have already occurred and does not support predictive control—for instance, prompting users to take breaks before quality declines become evident.

Given these limitations of existing approaches, it is highly valuable to develop methods that can estimate the likelihood of response quality degradation in advance, before the quality visibly deteriorates, and that enable proactive interventions.

B. Studies on Improving Response Quality

Several studies are focusing on improving response quality by enhancing user engagement and inducing behavior changes.

Sihang *et al.* [7] proposed a method to increase user engagement in crowdsourced microtasks by using a conversational interface instead of a traditional web-based interface. Zhang *et al.* [8] examined behavior change in the context of encouraging increased step counts and showed that the conversational style of information presentation—such as the level of detail or politeness—significantly influences the effectiveness of such interventions. These findings suggest that natural, interactive communication helps sustain motivation and attention, which also informs the adoption of conversational interfaces in our study. Other intervention studies have focused on maintaining users' engagement and concentration. Jeffrey *et al.* [9] reported that inserting appropriate breaks during long tasks can greatly improve user engagement. Similarly, Peng *et al.* [10] demonstrated that incorporating brief entertainment between crowdsourced microtasks can improve user engagement without sacrificing task performance. These strategies are particularly effective in mitigating fatigue and attentional decline.

Meanwhile, some studies have proposed interventions aimed at shaping user attitudes before or during task execution. Oyama *et al.* [11] proposed a method in participatory sensing where users express their commitment at the beginning of a task by tapping a button or shaking their device, thereby discouraging dishonest responses such as prioritizing speed over accuracy. Nakagawa *et al.* [12] proposed an interface using sliders and magnifiers to capture subtle user hesitation during responses through touch interaction logs. Mara *et al.* [1] proposed a method for detecting stress in workplace environments by extracting features from mouse and keyboard interactions as well as heart rate variability.

These studies focus on improving response quality by designing better annotation platforms, but they do not incorporate responsive (reactive) interventions to detect and address ongoing degradation. Therefore, detecting the signs of quality degradation in real time and implementing timely, personalized interventions is considered crucial for improving response quality through behavioral support.

III. Proposed Method

In this study, we focus on an annotation task in which participants evaluate the correctness of captions corresponding to images. The goal is to estimate the tendency of response quality degradation in real time during task execution. This section describes the method for collecting smartphone interaction data, which includes screen operations such as taps and scrolls, as well as sensor data such as device orientation and acceleration. It also outlines the approach for constructing a machine learning model to estimate quality degradation based on these features.

A. Overview of the Proposed Method

The assumed scenario for the annotation task and an overview of the proposed method are illustrated in Fig. 1. The

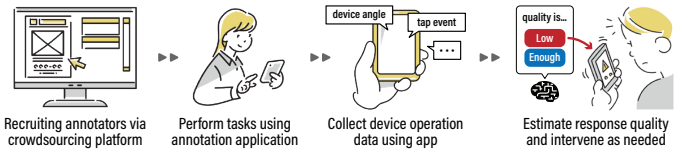

Fig. 1. Assumed scenario and proposed method

following describes each step in detail.

1) The requester of the annotation task recruits annotators via a crowdsourcing platform. In this study, the annotators are assumed to be general users who are native speakers of the annotation target language, rather than domain experts.
2) The annotators perform the requested annotation tasks using a dedicated application installed on their personal smartphones.
3) While performing the task, the application continuously and unobtrusively collects device operation data in the background. This includes screen interactions such as taps and scrolls, as well as sensor data such as device orientation and acceleration.
4) Features are extracted from the collected log data and input to a machine learning model, which then estimates the tendency of response quality degradation in real time. If the degradation is detected, the application intervene to the user for changing their behavior.

B. Overview of the Assumed Annotation Task

In this study, the target annotation task is the binary evaluation of the correctness of image captions. The annotation task is assumed to be performed using a custom smartphone application developed for this study.

In the task, a set consisting of an image and its corresponding caption is presented to the annotator. The annotator is required to determine whether the caption accurately describes the content of the image. If the caption is judged to be correct, the annotator selects "Yes"; otherwise, "No."

The detailed procedure of the annotation task is as follows:

1) The user logs in to the application using their ID and password.
2) From the task selection screen, the user selects the assigned task, which transitions to the annotation interface.
3) In the annotation interface, the task is presented through a conversational interface by a virtual agent.
4) The annotator reviews the presented image and the caption displayed below it. If the caption correctly describes the image, the annotator selects "Yes"; otherwise, they select "No."
5) Upon completing all assigned tasks, a completion message is displayed by the agent, and the annotation session ends.

C. Device Interaction Logging System and Feature Extraction

To capture device operation data during annotation tasks on smartphones, we developed a dedicated annotation application

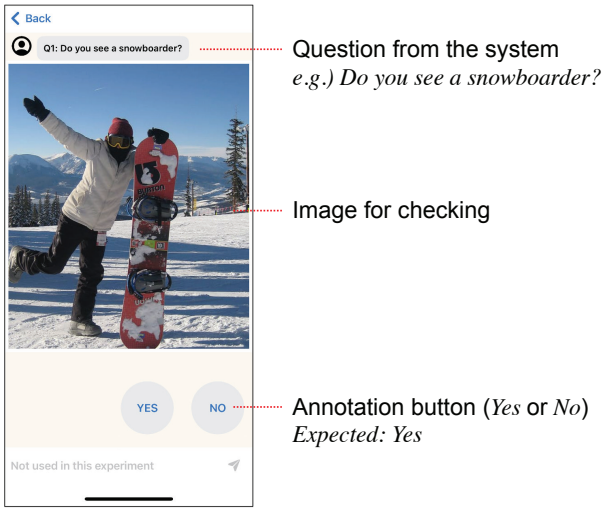

Fig. 2. Annotation application and assumed annotation task example

as shown in Fig. 2. In addition to presenting annotation tasks, the application continuously logs user interactions such as taps and scrolls, along with sensor data including device orientation and acceleration, in the background. These logs are stored in a local database for subsequent analysis.

The features extracted from the collected device interaction data are summarized in Table I. In addition to the features used in previous research by Gogami *et al.* [5], we introduced new features tailored to the operational characteristics of the binary image caption verification task used in this study. These include sensor-based features such as device orientation, as well as time-series-friendly features suited for real-time processing. The goal is to enable more flexible and accurate estimation of response quality degradation.

The target task—evaluating whether a caption correctly describes a given image—requires cognitive processing such as visual interpretation and logical judgment about the semantic match between image and caption. Due to these demands, annotators must engage in both perceptual and linguistic processing. When response quality declines, annotators may rush through tasks without completing these processes, resulting in unusually short response times or low accuracy. We assume such degradation manifests as measurable changes in time and accuracy. Thus, extracting features that reflect these variations is expected to support effective quality degradation estimation.

D. Response Quality Degradation Estimation Model

This section describes the method for constructing a machine learning model to estimate response quality degradation using the features derived from the device interaction data obtained via the application described in the previous section.

We adopt LightGBM, which was selected based on its superior performance in previous research by Gogami *et al.* [5], as the classification algorithm. Among the extracted features, some—such as screen coordinates and the number of taps—exhibit significant variation in scale across samples. Therefore, we apply standard score normalization (Z-score

TABLE I
EXTRACTED FEATURES

Feature	Unit
Timestamp of data submission	s
Task number	
Task ID	
Response time	s
Inactive duration	s
View position (y)	
Number of taps	count
Tap interval	s
Tap position (x, y)	
Number of scrolls	count
Scroll length	
Scroll duration	s
Scroll speed	
Orientation angle (x, y, z)	
Gyroscope acceleration (x, y, z)	
Acceleration (x, y, z)	
Gravity acceleration (x, y, z)	
Response (Yes / No)	
Response correctness	
Accuracy over previous 10 tasks	%

normalization) with mean 0 and standard deviation 1 to unify feature scales. To improve generalization performance and avoid overfitting, we optimize the model's hyperparameters using Optuna[1], an automated hyperparameter optimization framework.

IV. EXPERIMENT AND EVALUATION

A. Data Collection

First, we conducted data collection experiment using our application described in Section III, and built a dataset. This study was approved by the Ethics Committee for Research Involving Human Subjects at the Nara Institute of Science and Technology (Approval No.: 2020-I-2).

We recruited graduate students at the Nara Institute of Science and Technology (NAIST), with a total of 38 individuals participating in the experiment. Participants were asked to perform annotation tasks, during which the application continuously recorded device interaction data in the background. This dataset was used for training machine learning models. Each participant received a gift card worth 1,000 JPY as compensation upon successful completion of the experiment.

To mitigate bias caused by participants being aware of data collection, we did not inform them in advance that device interaction data would be recorded. Only task instructions were provided before the experiment. After the experiment concluded, participants were informed that device interaction data had been collected, and explicit consent for data usage was obtained at that time.

1) Overview of the Annotation Task: The annotation task used in the experiment is described here. Specifically, we used the GQA (Generative Question Answering) dataset [13], which consists of images paired with descriptive captions, as the basis for the annotation task.

[1]https://optuna.org/

Participants were asked to judge whether the caption accurately described the content of the corresponding image. If the caption was correct, they responded with "Yes"; otherwise, they responded with "No." The annotation procedure followed the steps outlined in Section III-B. In this experiment, 150 image-caption pairs were selected from the GQA dataset and presented to participants in a randomly shuffled order unique to each individual. The task session ended when either all annotation tasks had been completed or 30 minutes had elapsed from the start of the task, in which case the current task was completed and the session was terminated.

2) Dataset: This subsection describes the dataset obtained through the experiment. For each image-caption pair in the task set, a ground truth label indicating correctness was assigned in advance. The correctness of each participant's response was then evaluated by comparing it against this ground truth. Out of the 38 participants, we selected data from 30 participants who completed all tasks and provided consent for research use. The constructed dataset contains 2,015 samples of enough quality responses, and 2,159 samples of low-quality responses.

Since the annotation task used in this study is a binary-choice task, we defined a quality degradation state as a period in which the correct response rate fell below 40% over a sliding window of 10 consecutive tasks. This definition allows us to exclude temporary errors or random correct guesses, enabling the detection of meaningful trends in response quality degradation.

B. Model Evaluation Method

In this study, we evaluate the performance of the constructed models based on Precision, Recall, and F1-score from two perspectives: (1) within-individual prediction accuracy and (2) generalization performance across the participant population.

For within-individual prediction accuracy, we adopt an expanding time-series validation approach. For each participant, the data from an arbitrary task is used as the test set, and all preceding data in the time series is used as the training set. This time-based partitioning is repeated sequentially to construct a binary classification model. The prediction results from each fold are then used to evaluate the model's accuracy.

For generalization performance, we use data from a single participant as the test set and construct an ensemble model using the remaining participants' data. The prediction results obtained from this model are then evaluated. This procedure verifies whether the proposed method generalizes effectively to unseen users.

C. Experimental Results and Discussion

1) Within-Individual Model Construction Using Expanding Time-Series Validation: Fig. 3 shows the classification results of models constructed using expanding time-series validation for each participant, based on features extracted from device interaction data. We have confirmed this model achieved Precision of 0.723, Recall of 0.741, and F1-score of 0.731 to the label of "1 (low-quality)." These results suggest that

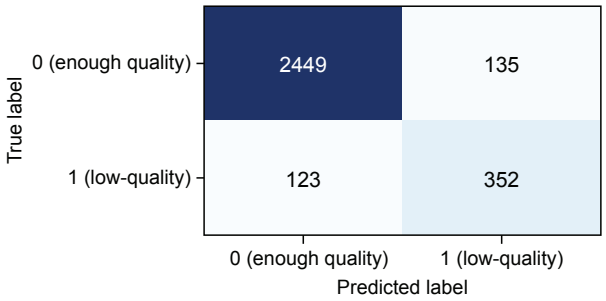

Fig. 3. Estimation result of within-individual model

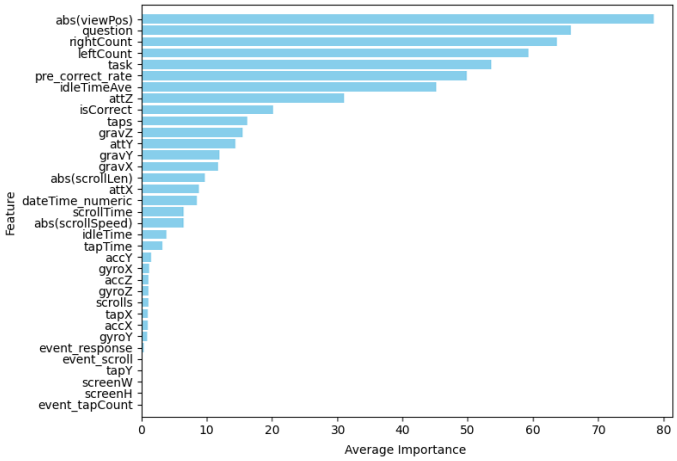

Fig. 4. Feature importance of within-individual model

the proposed approach is effective in estimating response quality degradation during annotation tasks using smartphone interaction data and has potential for practical deployment.

On the other hand, classification errors are observed in some cases. Given the binary nature of the task—determining the correctness of a caption associated with an image—there is a possibility that participants could occasionally produce correct responses even when not properly engaged in the task. Such instances result in a mismatch between the device interaction log and the correctness label, which may degrade model training and prediction performance.

Fig. 4 shows the top 20 features that most significantly contributed to classification performance across all participants. The vertical axis represents the names of the features, while the horizontal axis represents their importance scores—larger values indicate greater contribution to the classification results. From the figure, we observe that several features had particularly high importance: the absolute view y-position (`abs(viewPos)`), task content (`question`), the number of selections for each option (`rightCount`, `leftCount`), task progression index (`task`), and the accuracy rate of the previous questions (`pre_correct_rate`).

The prominence of screen position and task progression features suggests that user fatigue tends to accumulate as tasks progress, which may lead to a decline in response quality. Additionally, the importance of task-related features implies

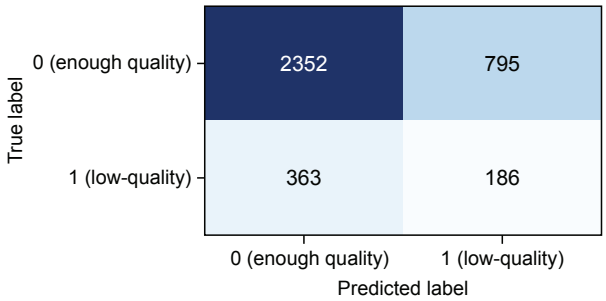

Fig. 5. Estimation result of LOPO model

that incorporating a quantitative assessment of task difficulty could further improve the accuracy of the classification model. The high importance of `pre_correct_rate` also indicates that modeling user reliability or tendencies over time (e.g., as a trust score) could be beneficial for predicting quality degradation. Moreover, the fact that the number of selections for specific options was a strong predictor suggests that participants may exhibit certain biases, such as favoring a particular option when uncertain, or may unconsciously select buttons that are easier to press when fatigued.

2) Evaluation of Generalization Performance with Ensemble Learning: Based on the individual models constructed in Section IV-C1, we conducted ensemble learning to evaluate the generalization performance of the proposed method. Fig. 5 shows the prediction results when using ensemble models trained on data from all participants except one, with the excluded participant's data used as the test set, i.e., Leave-One-Person-Out (LOPO) Cross-validation. We have confirmed this model achieved Precision of 0.190, Recall of 0.339, and F1-score of 0.243 to the label of "1 (low-quality)."

Compared to the results obtained from the intra-individual models using expanding time-series validation described in Section IV-C1, all evaluation metrics decreased significantly. One possible reason is that models trained on individual data learn patterns specific to each participant's device interaction and behavior. When applying ensemble learning, these individualized patterns may conflict or cancel out due to inter-participant variability, leading to degraded classification performance.

As also noted in the findings by Gogami *et al.* [5], intra-individual variability is often more informative than inter-individual variability when estimating response quality. The results of our study support this observation, suggesting that capturing personal behavioral trends is crucial for reliable quality estimation. Fig. 6, Fig. 7, and Fig. 8 illustrate the distributions of Precision, Recall, and F1-score obtained from the ensemble learning results. While examining these distributions, we observe the presence of outlier samples with notably low performance in both Precision and Recall.

One possible reason for these outliers is the presence of participants whose device interaction characteristics substantially differ from others, as shown in the feature clustering results

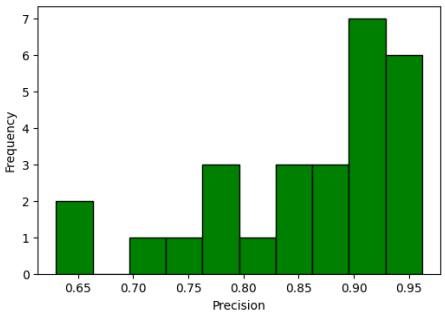

Fig. 6. Precision distribution

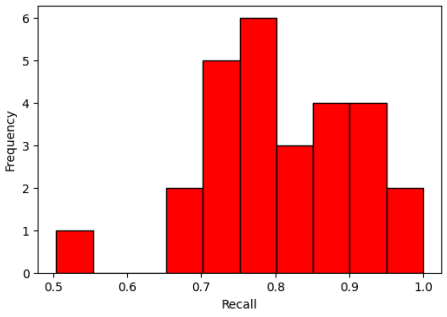

Fig. 7. Recall distribution

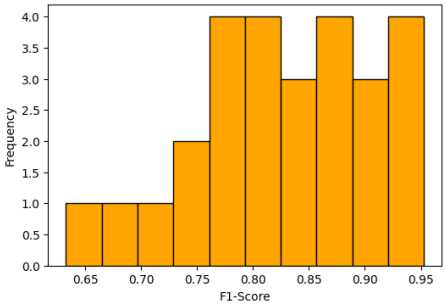

Fig. 8. F1-score distribution

discussed in Section IV-A2. The inclusion of such participants may have introduced greater variability into the overall model performance, thereby resulting in extreme values in specific evaluation metrics. To address this, increasing the sample size and conducting cross-validation only within participant clusters that exhibit statistically similar behavioral patterns may lead to improved model accuracy and robustness.

V. Conclusion

In this study, we proposed a method for real-time estimation of response quality degradation in crowdsourced microtasks, specifically focusing on binary evaluation of image-caption correctness. We constructed and evaluated a binary classification model using machine learning techniques. In evaluations using time-series validation on individual-level data collected in a university setting, the proposed model achieved an average Precision of 0.722, Recall of 0.741, and F1-score of 0.731. These results suggest that the proposed approach is potentially effective for estimating real-time response quality degradation in practical applications. However, in the generalization performance evaluation using ensemble learning, the model only achieved an average Precision of 0.190, Recall of 0.339, and F1-score of 0.243, indicating that the constructed model lacks sufficient generalization capability across participants.

For future work, we plan to improve classification accuracy by designing higher-level features that more precisely reflect user behavior based on device interaction data. Additionally, since the current experiment was conducted in a university setting, we aim to replicate the study in a more realistic crowdsourcing environment by collaborating with existing platform operators. Regarding the quality metrics for annotation tasks, we defined low-quality responses based on the correct response rate over sliding windows of ten tasks. In future work, we plan to incorporate standardized psychological measures, such as IMC and ARS used in social psychology, which would allow for direct comparison with prior studies. Moreover, while the current classification model focuses on binary classification using device operation data, the annotation application itself employs a conversational interface. Therefore, by integrating a natural language agent that provides feedback when quality degradation is detected, the model could be extended to multi-class classification that considers both the occurrence and causes of quality degradation.

References

[1] M. Naegelin, R. P. Weibel, J. I. Kerr, V. R. Schinazi, R. La Marca, F. von Wangenheim, C. Hoelscher, and A. Ferrario, "An interpretable machine learning approach to multimodal stress detection in a simulated office environment," *J. of Biomedical Informatics*, vol. 139, no. C, 2023.

[2] H. A. Simon, "Rational choice and the structure of the environment." *Psychological review*, vol. 63, no. 2, p. 129, 1956.

[3] D. M. Oppenheimer, T. Meyvis, and N. Davidenko, "Instructional manipulation checks: Detecting satisficing to increase statistical power," *Journal of experimental social psychology*, vol. 45, no. 4, pp. 867–872, 2009.

[4] M. R. Maniaci and R. D. Rogge, "Caring about carelessness: Participant inattention and its effects on research," *Journal of Research in Personality*, vol. 48, pp. 61–83, 2014.

[5] M. Gogami, Y. Matsuda, Y. Arakawa, and K. Yasumoto, "Detection of Careless Responses in Online Surveys Using Answering Behavior on Smartphone," *IEEE Access*, vol. 9, pp. 53 205–53 218, 2021.

[6] Y. Fukumitsu, Y. Matsuda, H. Suwa, and K. Yasumoto, "Detecting careless responses in dataset annotation using screen operation logs," in *2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom '24)*, 2024, pp. 775–780. [Online]. Available: https://doi.org/10.1109/PerComWorkshops59983.2024.10502811

[7] S. Qiu, U. Gadiraju, and A. Bozzon, "Improving Worker Engagement Through Conversational Microtask Crowdsourcing," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI'20, 2020, pp. 1–12.

[8] Z. Zhang, J. Miehle, Y. Matsuda, M. Fujimoto, Y. Arakawa, K. Yasumoto, and W. Minker, "Exploring the Impacts of Elaborateness and Indirectness in a Behavior Change Support System," *IEEE Access*, vol. 9, pp. 74 778–74 788, 2021.

[9] J. M. Rzeszotarski, E. Chi, P. Paritosh, and P. Dai, "Inserting micro-breaks into crowdsourcing workflows," in *The First AAAI Conference on Human Computation and Crowdsourcing*, ser. HCOMP'13, 2013, pp. 62–63.

[10] P. Dai, J. M. Rzeszotarski, P. Paritosh, and E. H. Chi, "And Now for Something Completely Different: Improving Crowdsourcing Workflows with Micro-Diversions," in *Proceeding of The 18th ACM Conference on Computer Supported Cooperative Work and Social Computing*, ser. CSCW'15, 2015, pp. 628–638.

[11] K. Oyama, Y. Matsuda, R. Yoshikawa, Y. Nakamura, H. Suwa, and K. Yasumoto, "A Method for Expressing Intention for Suppressing Careless Responses in Participatory Sensing," in *18th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, ser. MobiQuitous'21, 2021, pp. 769–782.

[12] T. Nakagawa, Y. Arakawa, and Y. Nakamura, "Augmented Web Survey with enhanced response UI for Touch-based Psychological State Estimation," in *2022 IEEE 4th Global Conference on Life Sciences and Technologies*, ser. LifeTech, 2022, pp. 91–95.

[13] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.