2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII) | 979-8-3315-1643-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/ACII63134.2024.00013

Incremental Multimodal Sentiment Analysis for HAIs Based on Multitask Active Learning with Interannotator Agreement

Thus Karjanapatchara*, Sixia Li*, Candy Olivia Mawalim*, Kazunori Komatani†, Shogo Okada*
*Japan Advanced Institute of Science and Technology, Nomi, Ishikawa, Japan
Email: {s2210402, lisixia, candylim, okada-s}@jaist.ac.jp
†Osaka University, Ibaraki, Osaka, Japan
Email: komatani@sanken.osaka-u.ac.jp

***Abstract*—Multimodal sentiment analysis (MSA) is critical in developing empathetic and adaptive multimodal dialogue systems or conversational agents that can naturally interact with users by recognizing sentiment and engagement. Addressing the challenges of collecting labeled data for MSA in human–agent interaction (HAI), this study introduces an innovative approach that combines active learning and multitask learning. Our efficient sentiment recognition model leverages active learning to select informative data for learning models, significantly reducing the labor-intensive data labeling process. Furthermore, we employ multitask learning to improve annotation (label) quality by evaluating alignment with true labels and interannotator agreement, thus enhancing the reliability of sentiment annotations. We evaluate the proposed multitask and active learning methods via a human–agent multimodal dialogue dataset that includes various types of sentiment annotations, which are publicly available. The experimental results demonstrate that by learning to predict the agreement score, multitask learning becomes better than single-task learning at capturing the uncertainties in the data. This study lays the groundwork for incremental learning strategies in MSA, aiming to adaptively understand user sentiments in human–agent interactions.**

***Index Terms*—Active learning, Multimodal sentiment analysis, Multitask learning**

I. Introduction

One of the ultimate goals of artificial social intelligence is to develop a multimodal dialogue system that can communicate naturally with users by recognizing sentiment and engagement statuses to generate empathic and adaptive responses. Even if large-scale language models [1] such as Chat-GPT[1] are publicly available, exploring how to implement such an empathetic and adaptive dialogue system is still the main challenge. A robust and accurate sentiment analysis model is vital for such a multimodal dialogue system.

However, specific challenges are encountered during the data collection process when working toward robust and accurate multimodal sentiment analysis (MSA) in human–agent dialogue interactions. First, collecting multimodal interaction data such as human-robot-agent interaction data is expensive, so it is still difficult to collect large-scale data in such a setting. An efficient multimodal learning methodology with low labeled data demands is needed. This leads to difficulties when training machine learning models with limited data. Second, although we need sufficient interaction data for a specific user to adapt the system to the specific interests of that user, we cannot always collect sufficient multimodal data (batch data) from unknown users. To solve the batch data limitation problem, the ability to obtain multimodal behavioral data with sentiment annotations is attained through dialogue interactions between the utilized system and the target user in an incremental manner. However, many MSA studies do not assume, that human–agent interactions are influenced by the data acquisition process; instead, they assume that a batch multimodal dataset is available for training machine learning-based MSA models.

In this study, we consider a scenario involving sequential learning strategies centered on active learning, which is a machine learning method. We assume that when the employed dialogue system interacts with a user and an utterance with an unclear sentiment level is observed from the user, the system can ask the user for their sentiment (ground truth) with respect to that utterance. As a result, first, the system can acquire pairs of multimodal features from the utterance and the sentiment labels (labeled data) observed in the utterance. Second, the system can incrementally update the sentiment recognition model by retraining it.

To satisfy the assumption made in this study, systems must possess three modules for (i) selecting uncertain utterances when the sentiment level of the user is unclear, (ii) annotating their true sentiment level for these uncertain utterances, and (iii) Conducting incremental learning on the acquired labeled samples composed of multimodal features from uncertain utterances and the corresponding true sentiment level.

We propose an efficient active learning strategy for MSA, in which (i) uncertain samples are estimated and (iii) retraining is performed using these estimated samples as training data. Here, there are two main methods for annotating the ground-truth sentiment label (ii). The first is to directly ask dialogue users about their true sentiment level. The second one is to ask multiple annotators to annotate the sentiment label of the user by showing the video of the dialogue after the dialogue

[1] https://chat.openai.com/

2156-8111/24/$31.00 ©2024 IEEE
DOI 10.1109/ACII63134.2024.00013

session. This study assumed the second method.

Active learning is effective for training models with small amounts of labeled data. Specifically, multimodal data obtained from interactions are sorted by uncertainty sampling; only uncertain samples (samples that are difficult to recognize) are transferred to annotators for labeling, so the sentiment recognition model can be efficiently trained by adding labeled data from such small uncertain samples.

In the early stage of active learning using small-sized training data, the uncertainty estimator is not sufficiently trained, so appropriately selecting samples is difficult. To overcome this issue, we utilize multitask learning (MTL) in the proposed active learning framework. Specifically, we set the learning tasks of predicting the true label and the agreement degrees of annotators (interannotator agreement (IAA) scores). The prediction of agreement degrees can reduce different annotators' subjective ambiguity and variability when annotating sentiment levels, thus improving the data quality and improving model performance. In addition to MTL, we utilize an ensemble of deep neural networks (DNNs) to estimate uncertain samples in a robust manner. MTL and deep ensembles (DEs) are frequently used and not novel, the novel; contribution is to show that an active learning strategy that integrates both MTL based on an annotator agreement recognition task and DE is efficient for MSA.

II. Related works

A. Sentiment analysis for spoken dialogue interactions

A computational method used to determine the sentiment or emotional tone expressed in a given set of data is called sentiment analysis, which is commonly referred to as opinion mining. It involves analyzing and classifying subjective information to determine whether it conveys a positive, negative, or neutral sentiment.

To learn about spoken dialogue interaction, we can observe the sentiment of the speaker in various ways. For example, we can infer the emotions of people with a conversation by listening to the tones of the voices that they use, which is part of the speech emotional task [2]. Speech emotion recognition (SER) is the process of extracting a speaker's inner turmoil from his or her speech. The pronunciation or spectrum aspects of the human voice are two ways in which its characteristics can be communicated [3] [4].

However, because everyone expresses emotions differently, it is difficult to tell which emotions are being conveyed on the basis of voice characteristics [5]. To overcome this complicated data situation, DNNs are used. DNNs outperform traditional methods in this field [6]. In addition to SER, DNNs also play a crucial role in multimodality scenarios. Deep learning-based techniques are the main trend in MSA [7]

Numerous previous investigations focused on developing multimodal machine-learning-based social signal identification models for human-robot-agent interactions. Multimodal processing at the dialogue and exchange level [8], [9], refers to the analysis and integration of multiple modalities (such as text, speech, gestures, facial expressions, and physiological signals) in the context of dialogue and exchanges between individuals.

Our research differentiates itself from other multimodal sentiment analysis (MSA) approaches by focusing on a scenario that employs sequential learning strategies centered on active learning. This is in contrast to existing methods that predominantly utilize supervised learning with batch training datasets.

B. Active learning and multitask learning

In real-world scenarios, acquiring data is a common problem in this field, and one aspect of this is limited labeled data. To overcome this problem, we use an active learning process. Active learning aims to select the most informative instance for annotation and can reduce the effort required for labeling.

The proposed uncertainty sampling strategy is one of the early methods used in active learning [10]. The purpose of this method is to measure uncertainty by using entropy or margins. It serves as a foundational technique for many active learning strategies. Another significant research area in active learning is the query-by-committee approach [11]. In this research, the author used an algorithm to perform querying according to the principle of maximal disagreement. Our approach is unique in that it integrates annotator agreement as a subtask within the multitask learning framework. This incorporation enhances generalization and effectively captures the impact of annotator consensus on the Multimodal Sentiment Analysis (MSA) process. The result showed that this algorithm led to a generalization error.

More recently, active learning techniques have been combined with a deep learning method [12] [13]. This research is related to the use of deep learning in the active learning process. The main purpose is to reduce the effort required from network security experts. Additionally, the combination of active learning with deep learning is used in the time-series field, which is a similar field to that of our work [14]–[16].

In the deep learning era, owing to its capacity to use shared representations via neural networks, MTL has drawn considerable interest [17]. In particular, in the computer vision field, various studies have been conducted on MTL [18]–[20], and social signal processing [21], [22]. One of the major challenges in MTL is the use of an effective shared representation that captures patterns across tasks. Zhang introduced a multitask DNN with a shared layer [23].

III. Data and Multimodal features

In this study, we used multimodal dialogue datasets: Hazumi1902 and Haumi1911 [24], which contain multimodal dialogue corpora between human participants and a virtual agent. The datasets are publicly available for academic research purposes [2]. The annotations focused on individual exchanges between humans and agents within conversations, capturing the sentiment levels and the desire to continue the conversation.

[2] https://www.nii.ac.jp/dsc/idr/rdata/Hazumi/

TABLE I
DISTRIBUTIONS OF CLASSES, SENTIMENT LEVELS (TSS), AND TOPIC CONTINUANCE LEVELS (TCS).

[%]	Hazumi 1902 [24]		Hazumi 1911 [24]	
Class	***TS***	***TC***	***TS***	***TC***
High	49.8	44.7	56.6	53.5
Neutral	42.7	39.4	36.1	29.2
Low	7.5	15.9	7.3	17.3
Total	**2,237 samples**		**2,439 samples**	

A. Recording settings

Hazumi includes multimodal data on conversations between participating users and a conversational virtual agent. The Wizard-of-Oz (WoZ) approach was used by a human operator to control the agent. To enable users to enjoy conversing with the agent, the operator chose his or her words properly and switched topics when needed. For example, if participants were not interested in the topic, the operator changed the topic. The operator played the part of a good listener if the participants seemed to enjoy speaking with the system. The participants included 10 males and 20 females, and their ages ranged from 20 to 70 years.

B. Sentiment annotations and statistics

Focusing on third-party annotations is crucial due to IAA, which enhances the robustness and reliability of sentiment analysis models. Self-reported sentiments are subjective and lack comparability, making it difficult to validate and improve model performance. Therefore, in this work, we focus on two labels per exchange: (1) third-party sentiment (TSs), i.e., sentiment levels annotated by third-party coders[3], and (2) topic continuance levels annotated by third-party coders (TCs). (1) The TS lies in the range from one to seven; one means that the user did not enjoy the corresponding exchange, and vice versa. (2) The TC lies in the range from one to seven; one means that the user did not want to continue the conversation, and vice versa. The TSs and TCs were annotated on 7-point scales; exchanges with values higher than 4.5 and lower than 3.5 were given high and low labels, respectively. The other exchanges were neutral. The label distributions are shown in Table I. The statistics show that the labels were mostly either ’neutral’ or ’high’, but the ’low’ class had the least number of labels in the data.

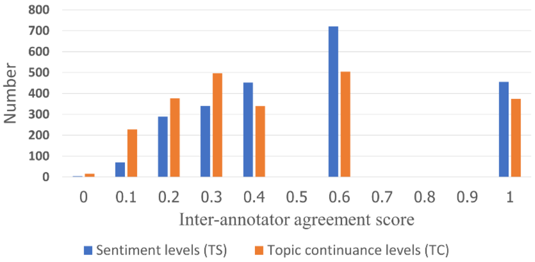

Fig. 1. Distributions of the IAA scores between sentiment level (TS) and topic continuance level (TC) in Hazumi1902

. We also focused on multitask learning, in which the auxiliary task was to predict the IAA score. The IAA score is a measure used to assess the level of agreement between two or more human annotators or raters who independently evaluate the same set of data. It is represented on a scale from 0 to 1, where 0 signifies no agreement among the annotators, whereas 1 indicates a high level of agreement. Fig. 1 shows the distributions of the IAA score at the sentiment level (TS) and topic continuance level (TC) in Hazumi 1902. The distributions revealed that sentiment levels tended to yield higher inter-annotator agreement (IAA) scores than did topic continuance levels (TCs). This might indicate that TCs have complicated properties for use in annotation tasks. To mitigate the issue of variability in labels provided by different annotators, we utilize the unweighted Cohen’s kappa statistic. This involves calculating the kappa coefficient for each pair of annotators, summing these values, and then computing the average (Equation (1): n_a denotes the number of annotators.). This method results in a lack of inter-annotator agreement (IAA) scores within the range of 0.7 to 0.9 in Fig. 1.

$$\mathrm{IAA} = \frac{\sum_{i=1}^{n_a-1}\sum_{j=i+1}^{n_a}\kappa_{ij}}{{}_{n_a}\mathrm{C}_k} \qquad (1)$$

C. Multimodal feature extraction

1) *Audio features.* Using the openSMILE toolkit [25], the acoustic characteristics were retrieved for each speech. The root mean square frame energy (RMSenergy), mel-frequency cepstral coefficients, zero-crossing rate, voice activity probability based on VAD (VoiceProb), and fundamental frequency (F0) were used as different types of acoustic parameters. Delta coefficients were also determined for these 94 characteristics.

2) *Linguistic features.* Linguistic features were extracted from the participants’ utterances. We used automatic-speech recognition and obtained text data. Then, we extracted the features contained in the text data. Following a morphological analysis using MeCab, the frequencies (also known as bags-of-words) and parts of speech of words were collected to produce 984 linguistic features.

3) *Visual features.* Regarding the visual features, we focused on the extraction of facial and motion features. By using OpenFace [26], we obtained ten facial feature points at the lips and both eyes. Calculations were performed to determine the absolute values of the between-frame accelerations and velocities. The incidence proportions of 18 action units were also employed. Microsoft Kinect V2 was used to obtain the joint locations of the participants along with their 3D coordinates. A total of 384 features were extracted using these positions.

[3] The dataset also has self-sentiment (SS), but we did not use it in this study

IV. Methods

We propose an efficient multimodal sentiment recognition model based on multitask active learning by adding labeled data from small uncertain samples while considering the agreement degrees of annotators (the reliability labels of their sentiment labels). An overview of the machine learning framework developed based on the proposed method is shown in Fig. 2.

A. Active learning

Active learning is a machine learning method where an algorithm proactively labels the most informative data points in an unlabeled set. The goal is to reduce the amount of labeled data needed while optimizing model performance. In the proposed active learning method, uncertainty sampling is used to choose the data points for labeling (Section IV-A1). Ensemble models enable us to accurately estimate uncertainty samples by aggregating the results output by multiple models, so we use an ensemble DNN as the base model. This method is called a DE (Section IV-A2).The procedure of the proposed active learning method with a DE is explained in Section IV-A3.

1) Uncertainty Sampling: In active learning, uncertainty sampling is the technique used to select informative and uncertain data points for labeling. Two uncertainty sampling methods are used in our experiments. First, the random sampling method randomly selects samples from each step and then retrains the constructed model with the selected samples. The second method is entropy-based sampling, which calculates the entropy of the predicted class probabilities. Entropy measures the uncertainty or randomness in a probability distribution. Samples with higher entropy values are considered more uncertain.

2) Deep Ensemble: DE is a machine learning technique that involves creating multiple deep learning models and combining their predictions to improve the overall performance and robustness of the resulting method. To obtain better results, we use a DE in terms of MTL and compare it with the single MTL strategy.

3) Procedure: In this study, active learning is conducted via the following process (Fig. 2):

1) **Training the model with initial training samples**: A machine learning model:$f(x)$ is trained with the initial training data. In this study, an ensemble of multiple neural network models with an MTL architecture is used as model $f(x)$.
2) **Transferring the data to an annotator**: Letting the unlabeled dataset be X_u, each sample in X_u is input to the trained model $f(x)$, and its uncertainty score is output from $f(x)$ (Section IV-A1). The top T samples with the highest uncertainty scores are transferred to the annotators.
3) **Labeling the uncertainty samples**: The oracle annotates the ground-truth sentiment labels (defined in Table I) for these T uncertainty samples.
4) **Retraining the model**: The initial model $f(x)$ is retrained with the T samples annotated by the oracle annotator.
5) **Classifying the data**: Once the training process is complete, DE MTL classifies the new unlabeled data. Drawing upon the principles of active learning, the classification model can be subject to periodic retraining, employing a diminished set of labeled data. Processes 1) to 5) are iterated.

B. Multitask learning

First, we extract the Hazumi dataset to form a feature vector (details in Section III). We then concatenate feature vectors from three sources to create a single input vector for the model. The MTL model, with two heads, performs prediction tasks and outputs the IAA score—a regression task measuring annotation reliability from below 0 to 1 (with scores below 0 indicating no agreement). This score is calculated using Fleiss' kappa value [27].

MTL trains a model on multiple related tasks simultaneously, enhancing generalization and performance. Our MTL framework trains task-specific models via a shared network architecture with three hidden layers (100 units) and two heads: one for predicting sentiments and another for IAA scores.

The first head predicts "High," "Neutral," or "Low" sentiments, while the second head, a regression task, predicts IAA scores based on these classifications. For example, to evaluate predicted sentiment levels (TSs), we calculate the IAA scores of the TSs.

1) Loss function of MTL: During the training process, we train the model with two tasks, the classification task of predicting sentiment levels or topic continuance levels, and the regression task of predicting agreement scores. We use the cross-entropy loss function (Equation (2)), which is the loss function that is typically used for classification tasks. Equation (2) shows the process of computing the cross-entropy loss function:

$$\mathcal{L}_{\text{CE}}(y, \hat{y}) = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_{i,j} \cdot \log(\hat{y}_{i,j}) \tag{2}$$

where L_{CE} is the cross-entropy loss, and y denotes the ground-truth labels, which form a one-to-one encoded vector with N samples and C classes, $\hat{y}$ is the predicted probability distribution, and $y_{i,j}$ is the ground-truth label for the i_{th} sample and j_{th} class, which is 1 if the sample belongs to class j_{th} and 0 otherwise.

Then, we use the mean squared error loss (Equation 3) for regression tasks, where the model predicts continuous values:

$$\mathcal{L}_{\text{MSE}}(y, \hat{y}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{3}$$

where L_{MSE} is the mean squared error loss, y is the ground-truth value for the i_{th} sample, $\hat{y}$ is the predicted value for the i_{th} sample, and N is the total number of samples.

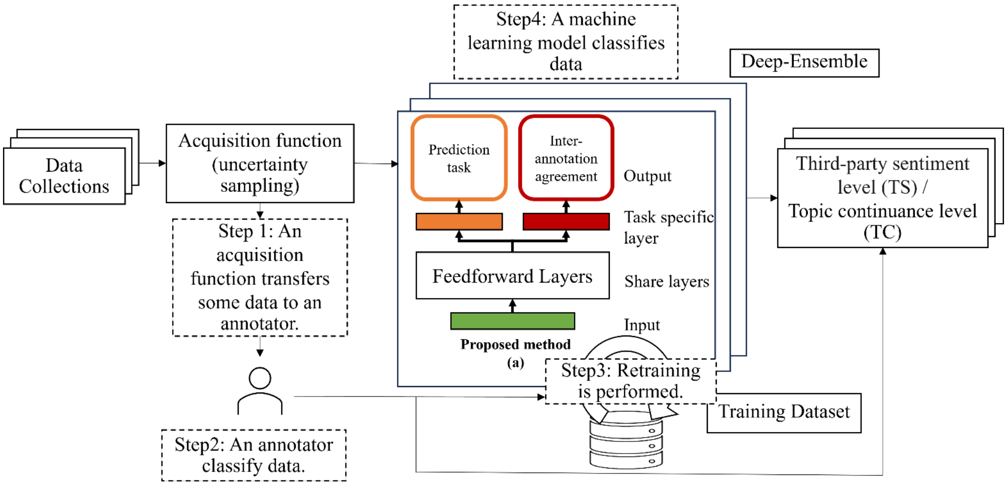
Step4: A machine learning model classifies data
Deep-Ensemble
Data Collections
Acquisition function (uncertainty sampling)
Prediction task
Inter-annotation agreement
Output
Task specific layer
Feedforward Layers
Share layers
Input
Proposed method (a)
Third-party sentiment level (TS) / Topic continuance level (TC)
Step 1: An acquisition function transfers some data to an annotator.
Step3: Retraining is performed.
Training Dataset
Step2: An annotator classify data.

Fig. 2. Overview of the proposed method based on an active learning model with (a) a DE and an MTL model

Finally, we compute the sum with the two tasks to compute The final loss for the MTL, is shown in Equation 4:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{MSE}} \tag{4}$$

V. Experimental settings

Our experiment focuses on a comparison between single-task learning and multitask learning in the active learning process. In this study, we start with 25 samples as the initial samples and conducts experiments by selecting 50 samples in each step via random sampling or uncertain sampling. We conduct an experiment to predict these labels and report the average accuracy achieved on the test set with 5-fold cross-validation.

This experiment consists of 37 steps conducted on the Hazumi 1902 dataset and 39 steps on the Hazumi 1911 dataset, and in each step, the models are trained for 100 epochs. The loss function for the classification task is the cross-entropy loss, and the IAA score is predicted, which we define as a regression task thus we use the mean squared error loss function. The learning rate in this experiment; is set to 0.001.

We utilize speech recognition to transcribe speech data into textual data, subsequently generating a bag-of-words vector. This vector is then concatenated with features derived from audio, linguistic, and visual modalities to form a unified feature vector, which is subsequently input into the active learning process.

During the training phase, the first step is to initially label the dataset to train the base model and then train the model using initial data. In this experiment, we use the initial data of 25 samples. Then, the trained model is applied to the samples. In this study, we initially start with 25 samples and calculate a measure of the uncertainty function for each sample. Afterward, we select a subset of the unlabeled data. The number of subset data is set to 50. Then, we label the selected subset data and feed them to the model for retraining. The selected data in our active learning framework are selected on the basis of their uncertainty scores derived from the current state of the model, not explicitly treated as time-series data. Finally, this step is repeated until the last step, and the model is evaluated on the test set after finishing each step.

1) Validation test and evaluation metric: We validate the 5-fold performance of the tested methods. We use the accuracy corresponding to the micro F measure as the evaluation metric (following previous work) via the Hazumi corpus [22]. The accuracy is calculated as the quotient obtained by dividing the number of correct class assignments by the total sample size.

As the primary evaluation metric, the total area under the accuracy plot (TAUP) for the active learning process is calculated according to the average accuracy achieved for all training sample sizes as the primary evaluation index. The TAUP obtained for the S-th step is calculated as follows:

$$TAUP_S = \frac{1}{S}\sum_{s=1}^{S} acc_s, \tag{5}$$

where s is the step index and acc_s denotes the accuracy achieved in step s. acc_1 denotes the accuracy of the model trained with the initial training set. During active learning process, evaluating the change in accuracy achieved by increasing the number of training samples chosen with uncertain sampling strategies is important. In particular, it is vital for the personality adaptive dialogue system to obtain a higher TAUP in an early step s. As a second strategy, the best accuracy is selected from all the steps of active learning. Additionally, we perform Fivefold cross-validation was performed to assess the performance and generalizability of the predictive model.

A. Comparative methods

In our experiment, the following eight models are implemented and compared to confirm the effectiveness of the proposed models.

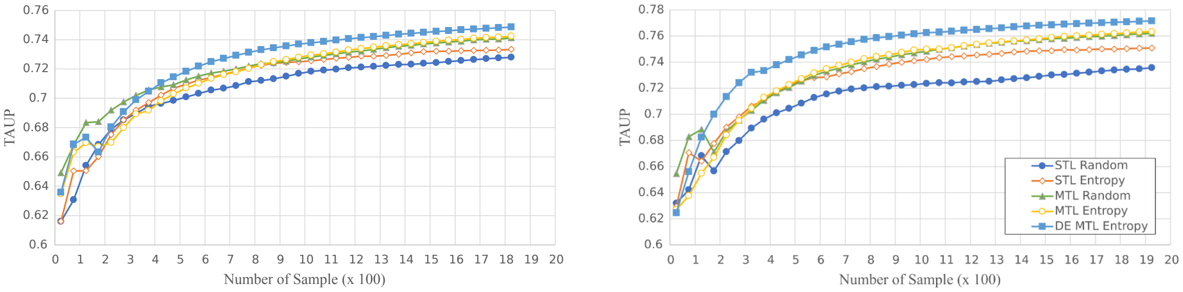

Fig. 3. The total areas under the plots ($TAUP_X$) produced by single-task learning (STL), multitask learning (MTL), and deep ensemble multitask learning (DE MTL) annotated at the sentiment level (TS) for Hazumi1902 and Hazumi1911.

- **STL + Random**: A DNN model is trained via single-task learning, and random sampling is applied as the uncertainty estimation function.
- **STL + Entropy**: A DNN model is trained via single-task learning, and entropy is applied as the uncertainty estimation function
- **MTL + Random**: A DNN model is trained via MTL, and random sampling is applied as the uncertainty estimation function.
- **MTL + Entropy (Proposed)**: A single neural network is trained via MTL, and entropy is applied as the uncertainty estimation function.
- **MTL + DE + Entropy (Proposed)**: Multiple DNN models are trained via MTL, and uncertainty estimation is performed via the aggregation of the outputs produced by multiple networks.

In our active learning experiment, we maintain consistent hyperparameters across trials to ensure a fair comparison, including fixed learning rate, batch size, and network architecture, with variability limited to the initial weight values.

VI. Results

Fig. 3 shows the changes in the total areas under the accuracy plots (TAUPs (Equation (5))) of the classification accuracies obtained by comparing the single-task learning and nonactive learning approaches (with random sampling) with the proposed multitask active learning method with a DE approach under an active learning process. The left and right panels of Fig. 3 show the trajectories of the TAUPs produced for the third sentiment prediction task conducted on Hazumi 1902 and 1911, respectively. In both subfigures, the proposed DE-based multitask active learning (MTL + DE + entropy) method yields the best results in almost all steps. These results, indicate that integrating uncertain sampling with DE and MTL is effective as an active learning strategy.

However, MTL + DE + Entropy does not obtain the best TAUP in the very early stage (with fewer than 37 training samples) of the active learning process, and the best result is obtained by MTL + Random in both cases. The main reason for this is that the uncertain samples are not accurately estimated by the initial model. Conversely, MTL + DE + entropy improves the accuracy and TAUP after training the model with a certain number of samples (more than 375 in Hazumi 1902, and more than 175 in Hazumi 1911). Here, the finding that MTL + random has the best TAUP means that although the uncertainty estimation process is not stable in the very early stage, MTL helps mitigate the degradation exhibited by the classification performance. The results show the effectiveness of performing MTL by optimizing the IAA score in the active learning task.

We summarize the TAUPs at the final step (after training on all samples), so we compare the proposed method with other baselines on the basis of their areas. Table II shows the area under the plot (AUP) results. The bold numbers indicate the best models for each label. As shown in the table, the proposed model (MTL + DE + entropy) achieves the best accuracies among all the models, with values of 74.9%, 71.3%, 77.2%, and 67.7% for Hazumi 1902 TS, Hazumi 1902 TC, Hazumi1911 TS, and Hazumi 1911 TC, respectively. Table III shows the best accuracy results derived from the active learning procedure. The bold numbers indicate the best model for each label. As shown in the table, the proposed model (MTL-DE + Entropy) outperforms the best baseline models by 73.5% and 78.8% on Hazumi 1902 TC and Hazumi 1911 TS, respectively. The performance of the proposed model is almost the same as that of the best baseline model for Hazumi 1911 TS and Hazumi 1911 TC. The primary metric is the AUP, so we discuss our results on the basis of Fig. 3 and Table II in the following sections.

A. Performance validation with statistical tests

We confirm that our proposed models do not improve the results merely by chance by implementing statistical hypothesis tests. Hypothesis testing is conducted to validate whether a significant difference in performance (AUP) is observed between the three baselines (STL + Random, STL + Entropy, and MTL + Random) and the proposed models (MTL + Entropy and MTL + DE + Entropy). For hypothesis testing purposes, we use the Wilcoxon signed-rank test to determine the statistical significance of our models. We perform a test on the cases that show the greatest AUP improvement when we adopt our multitask active learning methods. For a given

TABLE II
AREAS UNDER THE PLOTS OF SINGLE-TASK AND MULTITASK METHODS.

[%]	**Hazumi1902**		**Hazumi1911**	
	TS	***TC***	***TS***	***TC***
Baseline models				
STL + Random (†)	72.8 ± 2.7	68.5 ± 2.7	73.6 ± 3.2	65.4 ± 3.5
STL + Entropy (⋆)	73.3 ± 2.8	69.7 ± 4.1	75.1 ± 2.9	67.2 ± 5.2
MTL + Random (∗)	74.1 ± 2.5	70.8 ± 2.5	76.2 ± 3.4	64.9 ± 4.0
Proposed models				
MTL + Entropy	74.2 ± 3.1 †, ⋆	70.9 ± 2.7 †, ⋆	76.3 ± 3.5 †, ⋆	66.4 ± 4.4
MTL + DE + Entropy	**74.9** ± 3.2 †, ⋆, ∗	**71.3** ± 3.8 †, ⋆, ∗	**77.2** ± 3.0 †, ⋆, ∗	**67.7** ± 4.8 †, ∗

The specific symbols †, $\star, and$ ∗ denote the following baseline models: "STL + Random", "STL + Entropy", and "MTL + Random", respectively.

TABLE III
THE BEST ACCURACIES AMONG SINGLE-TASK AND MULTITASK METHODS.

[%]	**Hazumi1902**		**Hazumi1911**		**Average**
	TS	***TC***	***TS***	***TC***	
Baseline models					
STL + Random	75.4	72.1	76.7	69.0	73.3
STL + Entropy	76.0	72.7	77.5	**71.8**	74.5
MTL + Random	**77.0**	72.9	78.4	69.1	74.4
Proposed models					
MTL + Entropy	76.8	73.1	78.5	70.4	74.7
MTL + DE + Entropy	76.8	**73.5**	**78.8**	71.5	**75.2**

pair of models (three baselines and the two proposed methods in Table II), we determine whether there is a significant difference between the accuracies achieved for each total number of samples in the plot (e.g. Fig. 3).

From Table II, the best-proposed model (MTL + DE + Entropy) achieves significantly better accuracy than all the baselines on the Hazumi 1902 TS, Hazumi TC, and Hazumi 1911 TS datasets, and it attains significantly better accuracy than do the two baselines on the Hazumi 1911 TC dataset. These results demonstrate that by adding an active learning framework and MTL, the proposed method is more effective than the single-task learning method, which only optimizes the loss for the target sentiment labels and random sampling methods, which address various sentiment analysis tasks.

B. Effect of the DE multitask model

Our study aims to improve the performance of uncertainty sampling and the resulting classification accuracy by applying DE techniques for an MTL model. From Fig.3, the DE method exhibits remarkable consistency in its accuracy across different instances, suggesting its superiority over the alternative methods. From the hypothesis testing results shown in Table II, the AUP of the proposed DE method (MTL + DE + entropy) is slightly better than that of the non-DE method (MTL + entropy). The results show that the DE makes the active learning performance stable and accurate.

VII. Discussions

A. Advantages of MTL and improvements over the DE

Our results highlight the advantages of multitask learning (MTL) with annotator disagreement estimation over single-task learning in active learning. MTL nearly matches the performance of supervised learning, improving accuracy and data efficiency through shared knowledge and feature representations. Entropy sampling outperforms random sampling by selecting the most informative samples, enhancing learning efficiency.

Many multimodal sentiment analysis (MSA) studies assume a readily available dataset, overlooking the impact of data acquisition. Deep ensembles (DE) outperform single models by utilizing diverse predictions from multiple models trained on different data subsets or initializations. This diversity captures complex patterns, improving predictive performance and robustness. Notably, DE outperforms MLP during convergence, confirming the value of ensembles in the early learning stages.

B. Exploring active learning for self-sentiment labels

In the proposed approach, we assume that the ground truth label is aggregated of labels annotated by multiple annotators and the interannotator agreement (IAA) scores are available.

On the other hand, recognition of self-reported sentiment is an important task for a dialogue system to understand the user's real inner state [28] [29]. Exploring an active learning approach to recognize the self-sentiment state of the user, and asking for direct grand-truth labels of the uncertain samples during the process of dialogue interaction via the dialogue strategy of the system is a useful approach.

Future work will explore how to successfully ask a dialogue user about their true sentiment level. Of course, we need to find a way to ask whether the user enjoys the interaction, taking user privacy into consideration. In this scenario, the interannotator agreement (IAA) scores are not available and multitask learning is not available. In the future, we also plan to explore an efficient single-task active learning approach based on asking the dialogue user whether the user is enjoying the talking with the system, as well as multi-task active learning, in this paper.

VIII. Conclusions

In conclusion, our experimental investigation focuses on active learning strategies within the context of single-task and multitask learning frameworks. Our results consistently demonstrate the superiority of multitask learning (MTL) over single-task learning, highlighting the benefits of leveraging shared information across related tasks. Furthermore, our analysis comparing random sampling, and entropy sampling, reveals comparable performance, suggesting that entropy sampling methods more effectively capture informative instances for labeling than does the random sampling method. By implementing a deep ensemble (DE) in the experiments, the accuracy score is improved, which helps the individual models learn different patterns and relationships in the input data. Overall, our research enhances the understanding of active learning and MTL approaches, offering insights that can improve the development of effective and efficient learning methods across various domains.

Acknowledgement

This work was also partially supported by“Five-star Alliance”, the Cooperative Research Program of NJRC Materials and Devices, JSPS KAKENHI (22K21304, 22H04860, 22H00536, 23H03506), JST AIP Trilateral AI Research, Japan (JPMJCR20G6),

Ethical Impact Statement

This research did not involve the creation of new datasets; instead, it utilized a widely recognized public sentiment dataset that is commonly used in existing self-reported and third-party sentiment analysis studies. The dataset comprises data collected from individuals across various age groups and genders, ensuring that no personal identifying details are included. All participants were informed about how the dataset would be used and consented by signing a form. The dataset solely contains data that the participants consented to share publicly, including videos, texts, physical indicators, and labels, thereby addressing potential privacy issues with nonpublic information. The inconsistency among diverse age groups and sexes suggests the broad applicability of the findings of this study. However, the study’s focus on Japanese participants and the Japanese language limits its generalizability, necessitating further research across different cultural and linguistic contexts. We acknowledge that, like other sentiment analysis techniques, our system can potentially assess a person’s emotional state without their consent.

Therefore, we emphasize the importance of using these systems ethically and responsibly, adhering to ethical, moral, and legal standards. In the training session, the experiment was carried out on a single Nvidia RTX 3060 GPU. The training model required 30–60 minutes for the 5-fold validation experiment. This duration is considered acceptable and is not expected to produce a substantial carbon footprint.

References

[1] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, Z. Wu, D. Zhu, X. Li, N. Qiang, D. Shen, T. Liu, and B. Ge, “Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models,” 2023.

[2] K. Kaur and P. Singh, “Trends in speech emotion recognition: a comprehensive survey,” *Multimedia Tools and Applications*, pp. 1–45, 2023.

[3] S. Kuchibhotla, H. D. Vankayalapati, R. Vaddi, and K. R. Anne, “A comparative analysis of classifiers in emotion recognition through acoustic features,” *International Journal of Speech Technology*, vol. 17, pp. 401–408, 2014.

[4] I. Luengo, E. Navas, and I. Hernáez, “Feature analysis and evaluation for automatic emotion identification in speech,” *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 490–501, 2010.

[5] W. Lim, D. Jang, and T. Lee, “Speech emotion recognition using convolutional and recurrent neural networks,” in *Proc. Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2016, pp. 1–4.

[6] M. Chen, X. He, J. Yang, and H. Zhang, “3-d convolutional recurrent neural networks with attention model for speech emotion recognition,” *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[7] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, “Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research,” *IEEE Transactions on Affective Computing*, 2020.

[8] S. Katada, S. Okada, Y. Hirano, and K. Komatani, “Is she truly enjoying the conversation? analysis of physiological signals toward adaptive dialogue systems,” in *Proc. ACM International Conference on Multimodal Interaction (ICMI)*, 2020, pp. 315–323.

[9] W. Wei, S. Li, and S. Okada, “Investigating the relationship between dialogue and exchange-level impression,” in *Proc. ACM International Conference on Multimodal Interaction (ICMI)*, 2022, pp. 359–367.

[10] D. D. Lewis and W. A. Gale, “A sequential algorithm for training text classifiers,” in *Proc. International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. Berlin, Heidelberg: Springer-Verlag, 1994, p. 3–12.

[11] H. S. Seung, M. Opper, and H. Sompolinsky, “Query by committee,” in *Proc. Annual Workshop on Computational Learning Theory(SIGIR)*, 1992, pp. 287–294.

[12] I. Kansizoglou, L. Bampis, and A. Gasteratos, “An active learning paradigm for online audio-visual emotion recognition,” *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 756–768, 2019.

[13] H. Kawaguchi, Y. Nakatani, and S. Okada, “IDPS signature classification based on active learning with partial supervision from network security experts,” *IEEE Access*, vol. 10, pp. 105 713–105 725, 2022.

[14] C. Zimmer, M. Meister, and D. Nguyen-Tuong, “Safe active learning for time-series modeling with gaussian processes,” in *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.

[15] L. Pan, M. Kalander, Y. Zhang, and P. Wang, “Contrastive representation based active learning for time series,” in *Proc.IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*. IEEE, 2022, pp. 1–6.

[16] G. He, Y. Duan, Y. Li, T. Qian, J. He, and X. Jia, “Active learning for multivariate time series classification with positive unlabeled data,” in *Proc. IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2015, pp. 178–185.

[17] M. Crawshaw, “Multi-task learning with deep neural networks: A survey,” *arXiv preprint arXiv:2009.09796*, 2020.

[18] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, “Facial landmark detection by deep multi-task learning.” Springer, 2014, pp. 94–108.

[19] J. Dai, K. He, and J. Sun, “Instance-aware semantic segmentation via multi-task network cascades,” 2016, pp. 3150–3158.

[20] S. Liu, E. Johns, and A. J. Davison, “End-to-end multi-task learning with attention,” 2019, pp. 1871–1880.

[21] Y. Hirano, S. Okada, H. Nishimoto, and K. Komatani, “Multitask prediction of exchange-level annotations for multimodal dialogue systems,” in *Proc. ACM International Conference on Multimodal Interaction (ICMI)*, 2019, pp. 85–94.

[22] Y. Hirano, S. Okada, and K. Komatani, “Recognizing social signals with weakly supervised multitask learning for multimodal dialogue systems,” in *Proc. ACM International Conference on Multimodal Interaction (ICMI)*, 2021, pp. 141–149.

[23] Y. Zhang, Y. Liu, F. Weninger, and B. Schuller, “Multi-task deep neural network with shared hidden layers: Breaking down the wall between emotion representations,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 4990–4994.

[24] K. Komatani and S. Okada, “Multimodal human-agent dialogue corpus with annotations at utterance and dialogue levels,” in *Proc. International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2021, pp. 1–8.

[25] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor,” in *Proc. ACM International Conference on Multimedia*, 2010, pp. 1459–1462.

[26] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, “Openface 2.0: Facial behavior analysis toolkit,” in *Proc. IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2018, pp. 59–66.

[27] R. Artstein, “Inter-annotator agreement,” *Handbook of linguistic annotation*, pp. 297–313, 2017.

[28] S. Katada, S. Okada, and K. Komatani, “Effects of physiological signals in different types of multimodal sentiment estimation,” *IEEE Transactions on Affective Computing*, 2022.

[29] ——, “Transformer-based physiological feature learning for multimodal analysis of self-reported sentiment,” in *Proc. ACM International Conference on Multimodal Interaction (ICMI)*, 2022, p. 349–358.