Master's Thesis

Softlabel Classification for Multimodal Sentiment Estimation using Multiple Third-party Annotations

ZHAO ZHEN

Supervisor SHOGO OKADA

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

August 2024

Abstract

This thesis explores the challenges and proposes novel solutions in the field of multimodal sentiment analysis, with a particular focus on enhancing the accuracy of self-sentiment (SS) estimation in human-agent dialogue systems. The research addresses two critical issues in current sentiment analysis: the discrepancy between self-reported sentiments and third-party annotated sentiments, and the subjective nature of sentiment annotation leading to disagreements among annotators.

The study utilizes two shared multimodal dialogue datasets, Hazumi1902 and Hazumi1911, which contain rich multimodal data including linguistic, audio, and visual features. These datasets are unique in that they provide both self-reported sentiment labels (SS) and third-party annotated sentiment labels (TS), allowing for a comprehensive analysis of the differences between these two types of sentiment annotations.

A key contribution of this research is the development of a novel soft labeling approach for sentiment classification. This method addresses the inherent subjectivity in sentiment annotation by representing the sentiment as a probability distribution over possible classes, rather than as a single hard label. The soft labels are generated by considering the annotations from multiple third-party annotators, capturing the nuances and disagreements in their judgments.

The thesis presents a series of experiments that demonstrate the effectiveness of the proposed approach. Initially, a baseline model is established using a simple deep neural network (DNN) architecture that integrates audio, video, and text features. This baseline model is then compared with human-level performance, revealing a significant gap in accuracy.

To bridge this gap, the research explores several innovative strategies. First, it investigates the impact of using TS labels for samples where TS and SS labels are inconsistent. This approach shows a marked improvement over the baseline, highlighting the importance of addressing label inconsistencies. Building on this, the study introduces the soft label method, which proves to be even more effective. The soft label approach not only improves overall accuracy but also captures valuable information from minority annotators who may detect subtle emotional states that the majority miss. Furthermore, the research proposes a weighted loss function that assigns different importance to samples based on the consistency between their TS and SS labels. This technique further enhances the model's performance, bringing it closer to human-level accuracy.

This research explores approaches to address the challenge of estimating self-sentiment in dialogue systems, proposes methods for handling annotator disagreements, and investigates the potential of soft labels in representing the complex nature of human emotions. The findings of this study may have implications for the development of sentiment analysis systems, particularly in human-agent interaction contexts. By attempting to better capture the nuances of human emotion, these approaches could potentially contribute to improvements in human-computer interactions.

Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Background

In recent years, many spoken-dialogue robots and applications have been released. Most of these systems respond solely based on the text obtained through speech recognition. In contrast, humans interpret not only the content of words but also vocal tone, facial expressions, and body posture to understand the interlocutor's state. A dialogue system with these capabilities is called a multimodal dialogue system. Research on sensing users' mental states and obtaining valuable information is gaining attention under the name Social Signal Processing (SSP). SSP involves technologies for sensing information that manifests as users' internal states. Using machine learning with various sensor inputs, SSP can predict information such as "the user is currently showing interest in the topic of conversation."

Sentiment represents an emotional inclination, reflecting a tendency towards specific types of emotional experiences, such as positive or negative feelings [1]. Unlike emotions that are usually displayed externally, sentiment does not always manifest overtly [2]. The way emotions or sentiments are expressed is governed by emotional intelligence [3], influenced by personality traits [4], and dependent on contextual factors [5]. Thus, accurately determining the true sentiment within an individual poses a complex challenge that integrates aspects of psychology and social sciences.

Historically, sentiment analysis has predominantly utilized textual lexicon-based methods. However, with the advent of social media platforms incorporating not just text but also images and videos, the effectiveness of multimodal analysis has been extensively explored in recent years [2], [6]. This approach, known as multimodal sentiment analysis, combines both verbal and nonverbal information to assess sentiment [7], [8]. Textual, visual, and auditory features each offer unique attributes and complement each other in the process of sentiment analysis [7].

Assessing a user’s sentiment during a conversation is crucial for adaptive dialog systems. This assessment allows the system to dynamically adjust its conversation strategies to keep the dialogue engaging (i.e., real-time sentiment estimation). For example, if a user shows interest in a topic, the system should continue with that topic; if the user appears disinterested, the system should switch topics. This seemingly straightforward task has been addressed by numerous researchers through various methods [9]. Despite this, several challenges remain in sentiment analysis for adaptive dialog systems. Ideally, sentiment analysis would use self-sentiment (SS) labels provided directly by users. However, many studies rely on third-party sentiment (TS) labels, which do not always accurately reflect a user’s true feelings. Research by Truong et al. highlights discrepancies between SS and TS labels (observed emotion ratings) [10], [11], [12], [13].

Additionally, users may hide or alter their true emotions during conversations, complicating the estimation of their emotional states. While textual, audio, and visual data are valuable for estimating TS [8], models based on these observable signals may be less effective for SS estimation. Recent evidence suggests that physiological signals

can be useful for emotion recognition, indicating that they may help detect subtle negative sentiments that are not easily captured through acoustic or visual means.

1.2 Research Objective

The two issues previously mentioned are also connected to how ground-truth labels are determined. Typically, these labels are established by aggregating annotations from coders, often through methods like majority voting or simple averaging. However, annotating social signals, such as sentiment, is inherently subjective and ambiguous, leading to frequent disagreements among annotators and resulting in unreliable outcomes. Consequently, training samples with such unreliable labels can negatively impact the performance of supervised learning models. Multiple participants often provide inconsistent labels, meaning some annotators may have accurately detected the user's true emotions and thus assigned labels consistent with SS. However, if labels are obtained through aggregation, this important information may not be captured.

Therefore, this paper aims to develop a novel approach that effectively learns from the rich and diverse information provided by all annotators, rather than relying on simplistic aggregation methods. Specifically, this approach seeks to:

1. Capture and utilize the full spectrum of annotator perspectives, including minority opinions that may contain valuable insights.

2. Improve the accuracy of sentiment prediction models, particularly in estimating self-reported sentiments (SS) in human-agent dialogue contexts.

3. Address the challenge of unreliable labels by developing robust learning methods that can handle inconsistencies and disagreements among annotators.

4. Investigate methods to weight or prioritize different annotations based on their consistency with self-reported sentiments or other relevant factors.

By achieving these objectives, this research aims to significantly advance the field of sentiment analysis in human-agent interactions. The proposed methods could lead to more accurate and nuanced understanding of user emotions, potentially improving the responsiveness and empathy of dialogue systems. Furthermore, this research could have broader implications for machine learning approaches to other subjective and ambiguous tasks beyond sentiment analysis.

1.3 Thesis Outline

In Chapter 2, the existing work related to Multimodal Social Signal Processing and various methods for handling subjective labels are explored. Additionally, relevant work on soft labels associated with the proposed method is introduced.

Chapter 3 presents the datasets used in the research. The data preprocessing and feature extraction processes are explained, and the basic multimodal model is introduced.

In Chapter 4, based on the basic multimodal model from Chapter 3, an SS classification method based on soft labels is proposed. The soft label generation method is described.

Chapter 5 is dedicated to a comprehensive exploration of the methodology. The performance of the model is evaluated, and a detailed discussion of the results is provided. The implications of the findings are interpreted and explored.

The final chapter, Chapter 6, provides a conclusion of the research conducted, the outcomes achieved, and their broader significance.

Chapter 2
Related Works

2.1 Multimodal Social Signal Processing

Integrating multimedia or multimodal data, which includes audio, visual, and linguistic features, has proven to be a promising method for recognizing social signals such as emotions and engagement. This approach has been widely explored in human-robot and agent interactions, utilizing multimodal machine learning techniques [14]. Research in this field has frequently focused on analyzing multimodal behaviors to identify engagement levels [15, 16]. Recent innovations have led to the development of agent systems equipped with social signal sensing capabilities [17, 18], aimed at improving interpersonal communication skills. Additionally, methods for detecting user interests have been investigated [19, 20]. For example, Weber et al. [21] proposed a dynamic user modeling approach based on reinforcement learning to analyze reactions to a robot's jokes, while Nasihati et al. [22] designed dialogue management routines for multiparty interactions involving agents and infants.

Modeling social signals within multimedia contexts is a significant area of research. Biel et al. [23] introduced a multimodal analysis technique to predict personality impressions from YouTube videos, and Brilman et al. [24] developed a model to identify successful debaters using multimodal information. Recent studies have demonstrated the effectiveness of deep neural network (DNN) techniques in accurate multimodal modeling [25], particularly for social signal processing (SSP). Advanced methods such as temporally selective attention models [26], multi-attention recurrent networks [27], memory fusion networks [28], and tensor fusion networks [29] have been proposed for multimodal sentiment analysis. Furthermore, group detection during natural social interactions based on standing conversations has been studied using ubiquitous and multimodal sensing technologies [30].

While traditional research has focused on single annotation labels, such as engagement, communication skills, personality, or humor, employing multiple labels for the same dataset can enhance model accuracy. Hirano et al. [31] introduced a multimodal modeling approach with multitask learning to identify various labels, including interest levels, sentiment levels, and next-action decisions, facilitating adaptive strategies in multimodal dialogue systems. Despite attempts to aggregate results from multiple coders or self-reported annotations to establish ground truth labels, coder disagreement continues to impact SSP accuracy. To address this issue, Hirano et al. [32] proposed a weakly supervised learning (WSL) algorithm, which allows for the development of robust SSP models even when dealing with inaccurate labels from human-system dialogue interactions. This approach highlights the potential of WSL to overcome the limitations of traditional annotation methods and improve the reliability of social signal recognition.

2.2 Machine Learning with Subjective Annotations

Social signal perception is inherently subjective, often leading to disagreements among annotators, which can complicate efforts to improve model accuracy. Addressing this challenge involves two primary approaches. The first approach focuses

on defining reliable labels from subjectively annotated data, while the second approach integrates discrepancies among labels provided by multiple coders [33, 34, 35].

Previous research has explored various methods to manage coder disagreement, frequently employing techniques such as majority voting to merge labels from multiple coders [36]. For instance, Ozkan et al. [35] developed a two-step conditional random field (CRF) model specifically for predicting backchannels, addressing inconsistencies in coder annotations. Inoue et al. [33] introduced a hierarchical Bayesian model designed to evaluate user engagement in human-robot interactions, estimating both engagement levels and coder characteristics as latent variables. Kumano et al. [34] proposed a probabilistic model to integrate labels of perceived empathy, focusing on the co-occurrence of gazes and facial expressions between participants. Lotfian and Busso [37] recommended a machine learning curriculum for emotion recognition in speech, aimed at improving deep neural network (DNN) training efficiency by accounting for disagreements in crowdsourced labels. Additionally, Hirano et al. [32] put forward a weakly supervised learning (WSL) strategy to reduce the impact of unreliable annotated labels on training datasets.

Each of these approaches provides a distinct strategy for addressing the challenges associated with subjective annotation. By defining reliable labels or incorporating differences among coders, these methods enhance model robustness. Integrating these strategies into the training process can significantly improve the performance of models in social signal processing.

2.3 Machine Learning with Softlabel

Current research typically addresses annotator disagreement in two main ways: capturing the diversity of annotators' beliefs or assuming the existence of a single ground truth label despite the disagreement [38]. Aggregating annotator disagreement usually involves two approaches: converting labels into a one-hot hard label [40] or modeling disagreement as a probability distribution with soft labels [41].

Wu et al., building on Collins et al. [42], examine how soft labels can be generated from a small group of annotators by incorporating additional information, such as their self-reported confidence. This approach offers advantages over traditional hard or soft label aggregation methods, which often require extensive annotator resources and may depend on potentially unreliable crowd-sourced inputs [43]. Wu et al.'s focus is on contexts where the latter approach—assuming a single ground truth—is more appropriate. Consequently, they emphasize traditional evaluation metrics like the F1-score, which depend on a gold-standard label, despite the rise of alternative evaluation methods that do not rely on aggregated labels [39]. Their methodology is feasible due to their use of high-agreement test sets, where the 'true' label is relatively well-defined.

2.4 Characteristic of this Study

A key innovation is the development of a novel soft labeling method in this study, which represents sentiment as a probability distribution. This approach effectively addresses the inherent subjectivity in sentiment annotation, allowing for a more nuanced representation of emotional states. Another significant contribution is the introduction of a weighted loss function strategy. This strategy assigns varying importance to

samples based on the consistency between their TS and SS labels, enhancing the model's ability to learn from challenging cases and improving overall performance. The study employs specialized datasets, specifically the Hazumi1902 and Hazumi1911 datasets, which are designed to analyze adaptive dialogue strategies aimed at increasing user engagement. These datasets enable a focused evaluation of sentiment analysis techniques in the context of human-agent interactions. Rigorous performance benchmarking against human-level accuracy is a cornerstone of this research. The proposed methods are evaluated to provide a clear benchmark for their effectiveness, ensuring that the developed techniques meet high standards of performance. By integrating multiple annotation perspectives, the study captures a more comprehensive view of sentiment. Considering annotations from various third-party annotators allows for the inclusion of minority opinions and detection of subtle emotional states that might otherwise be overlooked. The research is particularly oriented towards improving sentiment analysis in human-agent interactions, with implications for developing more empathetic and responsive dialogue systems.

Chapter 3
Method

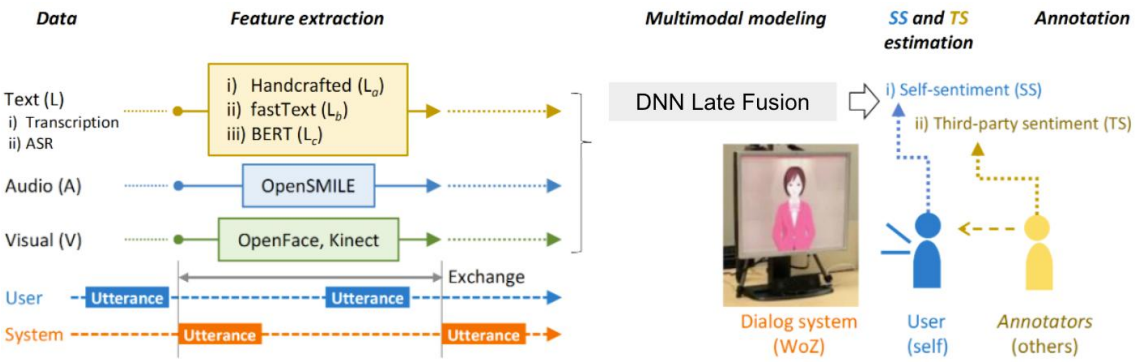

Figure1: Overview of the Estimation of Self-sentiment and Third-party Sentiment at the Exchange Level [32]

3.1 Data

The utilized framework is depicted in Figure 1. This framework is frequently employed in multimodal sentiment analysis [32]. Two shared multimodal dialogue datasets, Hazumi1902 and Hazumi1911, were used, both of which are part of the corpus described in [44].

3.1.1.Datasets Description

A virtual agent, MMD-agent, was used as the interface, as shown on the lower right-hand side of Figure 1. Participants interacted with the agent via a display. Their behaviors were recorded using a video camera and a Microsoft Kinect V2 sensor. The virtual agent was manipulated using the WoZ method, where the operator pretended to be a dialogue system. There was no specific task; instead, chat dialogues were used. The Hazumi1902 and Hazumi1911 datasets were collected to analyze an adaptive dialogue strategy aimed at enhancing user engagement. The operator selected utterances and appropriately changed topics. For instance, if participants showed disinterest in a topic, the operator switched topics. Conversely, if participants seemed to enjoy the conversation, the operator acted as an attentive listener. Due to issues encountered during data collection, the final number of participant data available for the experiment was: 28 participants (8 males and 20 females) for Hazumi1902, and 26 participants (12 males and 14 females) for Hazumi1911. Their ages ranged from 20 to 70 years. The

recording settings for Hazumi1902 were the same as those for Hazumi1911.

3.1.2.Labels Definitions and Statistics

In this experiment, the following two labels will be used: (1) self-sentiment levels felt by participants (SS), and (2) sentiment levels annotated by third-party coders (TS). Both SS and TS assist the system in recognizing whether the user is enjoying the dialogue and adapting its utterances accordingly.

Averaged annotated scores (Aas) were computed for each sample and subsequently transformed into ternary labels (high, neutral, and low). For the (1) SS and (2) TS, which were annotated using 7-point scales, exchanges with Aas values above 4.5 were categorized as high, and those below 3.5 were categorized as low. All other exchanges were classified as neutral. The distribution of these labels is presented in Table 1.

Table 1: Label Distribution (%)

	Hazumi1902		Hazumi1901	
Class	(1)SS	(2)TS (Aas)	(1)SS	(2)TS (Aas)
High	49.1	49.8	45.3	56.6
Neutral	30.5	42.7	34.8	36.1
Low	20.4	7.5	19.9	7.3
Total	2,337 samples		2,439 samples	

3.2 Multimodal Features

The multimodal feature set was extracted in the same manner as in [32].

3.2.1.Linguistic Feature

Linguistic feature sets for constructing multimodal models were derived from sentence representations using BERT [45], a language representation model renowned for its state-of-the-art performance across various NLP tasks. Effective pretraining of language models is crucial for achieving high performance [45][46]. Recently, a pretrained Japanese BERT model was developed at Tohoku University, demonstrating superior results compared to traditional bag-of-words models in tweet emotion recognition [47]. This study employed the pretrained Tohoku BERT model. Within the BERT framework, utterances from participants and systems in each exchange were separated by a special token ([SEP]). The sequences were tokenized using MeCab and segmented into subwords through the WordPiece algorithm. Activations from the second-to-last hidden layer of the BERT model were then averaged, producing a single vector of length 768 [45]. This vector was used as the input feature vector for each model, simplifying the process of feature extraction and facilitating the integration of additional modalities.

3.2.2.Audio and Visual Feature

Audio features were extracted from the participants' speech using openSMILE*12, employing the INTERSPEECH 2009 Emotion Challenge feature set (IS09) [48], resulting in a total of 384 dimensions. From webcam images, two types of facial expression features were extracted using OpenFace [49]: landmark features and action unit features, totaling 66 dimensions. The extraction process for the 48-dimensional landmark features involved obtaining two-dimensional coordinates of 12 facial landmarks around the eyes, mouth, and eyebrows. For each of these points, four statistical measures were calculated: the maximum value of absolute velocity, mean value, standard deviation, and maximum value of absolute acceleration. The 18-dimensional action unit features were derived from a pre-trained model in OpenFace, which detects the presence of 18 types of action units in each frame. The proportion of frames in which each action unit was detected during the conversation was used as features. Additionally, motion data from the hands, shoulders, and head were recorded using a Microsoft Kinect sensor [50], with calculated speed and acceleration serving as motion features. In total, 86 visual features were derived from facial expressions and motion activities. The data was standardized using Z-score normalization.

3.3 DNN Late Fusion

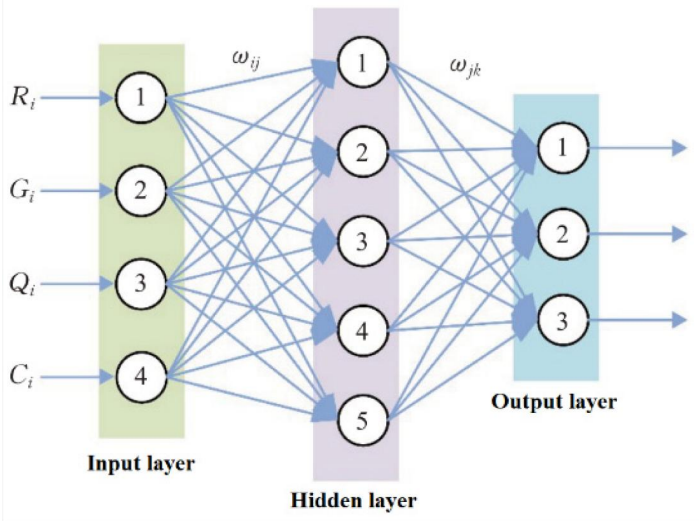

Figure 2: Feed-forward Fully Connected Neural Network

A simple neural network in Figure 2. also known as a single-layer neural network, consists of an input layer, one hidden layer, and an output layer. The input layer receives and formats external data for processing. The hidden layer contains neurons that learn and extract features from the input data, with the output computed using an activation function. The output layer generates the final prediction, with the number of neurons typically matching the number of categories in the task-one for binary classification and one per class for multi-class problems. Training is achieved through the backpropagation algorithm shown on Figure 3, which adjusts the network's parameters to minimize the difference between predicted and true results by iteratively

computing and reducing the loss function until an acceptable level is reached or the training iterations are complete.

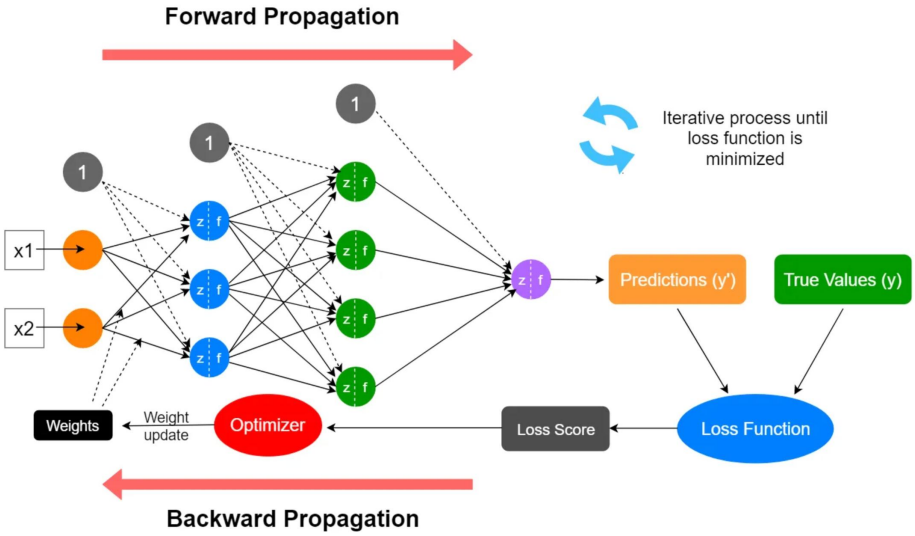

Figure 3: Backpropagation

In the context of more complex scenarios, such as multimodal learning, a different approach called late fusion comes into play Figure 4. Late fusion is a technique where different modalities are processed separately through their dedicated models, each extracting features or generating predictions independently. The outputs from these models are then combined at a later stage, just before making the final decision. This fusion can involve concatenating features, averaging predictions, or employing more sophisticated methods like multimodal neural networks. By allowing each modality to be processed in the most suitable way before integration, late fusion facilitates more nuanced and effective decision-making.

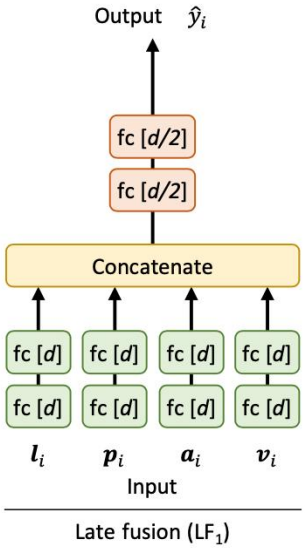

Figure 4: DNN Late Fusion

In this experiment, a multimodal neural network classifier was used, designed to integrate audio, video, and text features for classification tasks. The architecture consists of separate fully connected layers for each modality: audio, video, and text. Each of these layers projects the respective input features into a common hidden dimension space. Specifically, the model has three input layers corresponding to the three modalities: audio with an input dimension of 384, video with an input dimension of 86, and text with an input dimension of 768. These inputs are each passed through a fully connected layer with a hidden dimension of 32, followed by a ReLU activation function. To prevent overfitting, a dropout layer with a dropout probability of 0.5 is applied after the concatenation of the modality-specific outputs. The concatenated feature vector is then fed into a final fully connected layer that maps it to the output layer, which has three units corresponding to the number of classes. The model was trained using a batch size of 256 and a learning rate of 0.002 over 10 epochs. The architecture and training procedure effectively leveraged the multimodal inputs to enhance the classification performance.

Chapter 4
Proposed Model

The model and feature processing parts of this method are consistent with previous research, but a new perspective is proposed in the learning process. As previously noted, discrepancies among annotators often occur when labeling emotions. This method accommodates these differences by treating all third-party labels as soft labels, represented as a probability distribution. Additionally, third-party labels are used to estimate the self-reported labels. The data is divided into two groups based on whether the third-party labels agree with the self-reported labels, and different weights are applied to the loss calculation accordingly.

4.1 Process of Generating Softlabel

To account for discrepancies among annotators when labeling sentiment, the following method is used to generate soft labels from the annotations. This method accommodates an arbitrary number of annotators and label categories. The steps are as follows:

1. Collect Annotations:
   For each sample, collect annotations from multiple annotators. Each annotator labels the sample with one of the possible emotion categories.

2. Count Frequencies:
   For each sample, count the frequency of each label. Let N be the total number of annotators, and let fi be the frequency of the i-th label.

3. Calculate Probability Distribution:
   Convert the frequencies into probabilities by dividing each frequency by the total number of annotations. For a sample with C possible labels, the probability Pi for the i-th label is calculated as:

$$P_i = \frac{f_i}{N}$$

4. Generate Soft Label:
   Use these probabilities to create a soft label vector. The soft label for a sample will be a C-dimensional vector where each element represents the probability of the corresponding label.

For example, consider a scenario with three possible emotion categories (low, middle, high) and five annotators. Suppose the annotations for a sample are low, low, middle, high, and low. The steps are as follows on Figure 5.

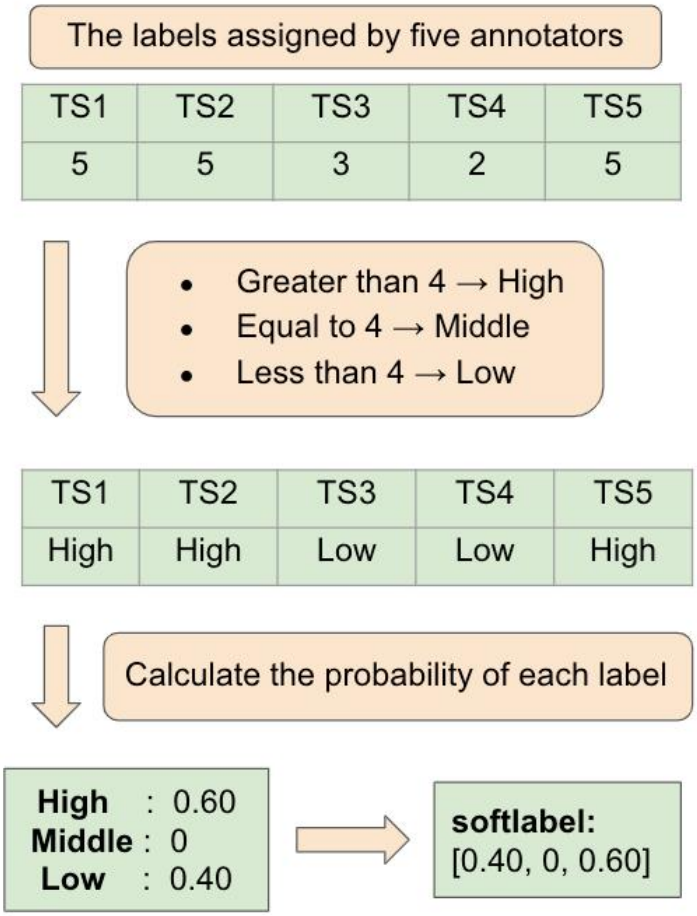

Figure 5: Example of Generate Softlabel

4.2 Train with Softlabel

For soft label learning, Softmax Cross Entropy is used, defined as follows:

$$L = -\sum_{k=1}^{K} q_k \log p_k$$

where q is the true probability distribution of the labels, p is the predicted probability distribution, and C is the number of classes. The Softmax function is used to convert the model's raw outputs into a probability distribution, and the cross-entropy loss measures the divergence between this predicted distribution and the true distribution.

In the context of soft label learning, the data is categorized into two groups based on the agreement between third-party labels and self-reported labels. The loss calculation is then weighted according to this categorization. The weighting scheme is defined as follows:

Loss for Consistent Labels (L_c):

This loss is calculated for samples where the third-party labels and the self-reported labels are consistent (i.e., they agree). The loss for these samples is computed using the Softmax Cross Entropy function and can be assigned a specific weight to reflect the reliability of these labels.

Loss for Inconsistent Labels (L_{inc})**:**

This loss is calculated for samples where the third-party labels and the self-reported labels are inconsistent (i.e., they disagree). This loss may be weighted differently or reduced to account for the uncertainty associated with these labels.

The total loss L is a weighted sum of the losses for consistent and inconsistent samples:

$$loss = L_c + \alpha L_{inc}$$

where α is the weight applied to the losses for consistent and inconsistent samples, respectively.This approach allows the model to prioritize samples with more reliable labels (consistent labels) and mitigate the impact of potentially noisy labels (inconsistent labels), leading to more robust training and better overall performance.

Chapter 5
Experiments

5.1 Evaluation Procedure

To evaluate the models, a cross-validation method (leave-one-person-out cross-validation, LOPOCV) was applied. In LOPOCV, the samples corresponding to each exchange between one participant and the dialog system were used as the test data, and the remaining samples were used as the training data. This procedure ensured that the test data from one participant were completely excluded from the training dataset, thereby avoiding overestimation. The accuracy and macro F1-score (F1) were calculated for each evaluation. F1 is particularly useful for imbalanced datasets. The F1 score is the harmonic mean of precision and recall. Here are the steps to explain the F1 score using a confusion matrix shown on Figure 6. Which includes four key quantities:

True Positives (TP): The number of instances correctly predicted as positive.
False Positives (FP): The number of instances incorrectly predicted as positive.
True Negatives (TN): The number of instances correctly predicted as negative.
False Negatives (FN): The number of instances incorrectly predicted as negative.

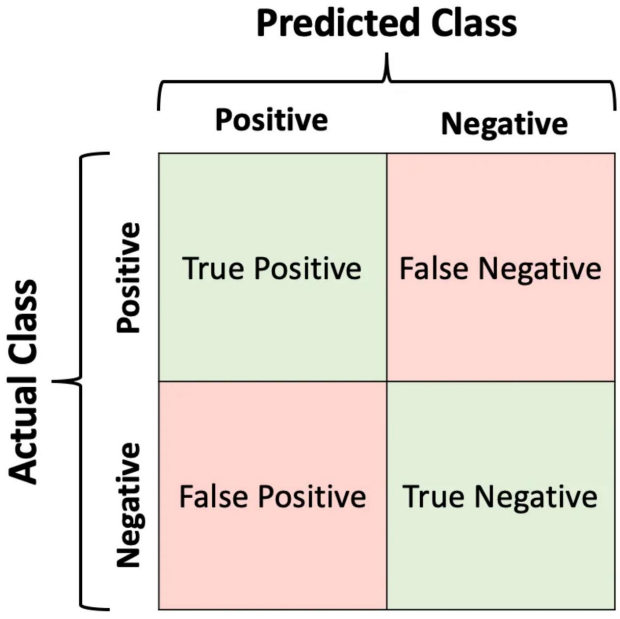

Figure 6: Confusion Matrices

Precision measures the proportion of correctly predicted positive instances out of all instances predicted as positive. It is calculated as:

$$Precison = \frac{TP}{TP + FP}$$

Recall measures the proportion of actual positive instances that were correctly predicted. It is calculated as:

$$Recall = \frac{TP}{TP + FN}$$

The F1 score is the harmonic mean of precision and recall. It provides a single metric that balances both precision and recall. The formula is:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

The average accurcators for the binary classification tasks. All experiments were performed three times with random initialization, and the evaluation values were calculated as averages across the three repetitions. These evaluation values were then compared among the models based on each modality or combination of modalities.

5.2 Baseline and Accuracy of Human Level

First, the goal is to establish a baseline for the model by directly classifying SS using the aforementioned model, with the results serving as the baseline. Additionally, to compare with human-level estimates, SS is evaluated against the human-annotated labels, TS (Aas), to obtain a 'human-level' reference. The results are summarized in Table 2. It can be observed that the accuracy of the model baseline is considerably lower than the human-level performance. The reasons for this disparity will now be analyzed.

Table 2: SS Estimation Baseline and Human Performance

	Hazumi1902		Hazumi1901	
	baseline	human	baseline	human
acc	0.4688	0.5383	0.4399	0.5433
f1	0.3372	0.4191	0.3298	0.4119

5.3 Experiments with TS Hardlabel

In a dialogue, participants may exhibit emotional states that do not match their true feelings, leading to discrepancies between self-reported and third-party observed labels. For example, a person may appear outwardly happy but actually feel sad internally. During the learning process of an SS classification model, samples that appear "outwardly happy" should have relatively similar features. However, if these samples simultaneously have opposing labels, such as "happy" (High) and "unhappy" (Low), this could lead to unstable or ambiguous decision boundaries, affecting classification performance. Next, the verification will be conducted to determine whether the low accuracy in SS classification is attributable to samples with inconsistent TS/SS labels.

First, the number of samples with consistent and inconsistent TS/SS labels will be visualized. It will be observed that nearly half of the total samples in both the Hazumi1902 and Hazumi1911 datasets have inconsistent TS/SS labels on Table 3.

Table 3: Consistent/Inconsistent Label distribution (%)

	Hazumi1902	Hazumi1911
consistent	1,261	1,294
inconsistent	1,076	1,145
Total	2,337	2,439

The data is divided into two groups based on the consistency between third-party labels (TS) and self-reported labels (SS): consistent data (Data(c)) and inconsistent data (Data(inc)). The SS classification model is then trained separately on these two groups. The results, as shown in the table, align with expectations. Specifically, the accuracy of Data(inc) is very low, demonstrating that samples with inconsistent TS/SS labels negatively impact the model's performance. To address this issue, the SS labels for Data(inc) are directly replaced with the TS(Aas) labels to ensure label consistency. This modified dataset is then combined with Data(c) for training the SS classification model. This adjustment results in a significant improvement in overall accuracy compared to the baseline. The results are shown in Table 4.

Table 4: Experiments with TS Hardlabel

	Hazumi1902			Hazumi1901		
	Data(c)	Data(inc)	all with TS(Aas)	Data(c)	Data(inc)	all with TS(Aas)
acc	0.5342	0.2342	**0.5250**	0.5144	0.2505	**0.5239**
f1	0.3825	0.1890	**0.3855**	0.3918	0.2095	**0.3852**

5.4 Experiments with TS Softlabel

In the previous subsection, it was observed that using TS(Aas) labels for samples with inconsistent TS(Aas) and SS labels enhanced the accuracy of SS label prediction. While TS(Aas) represents the average level of third-party labels, discrepancies may arise due to the subjective nature of annotator judgments. The next step is to analyze the relationship between third-party labels provided by different annotators and SS labels.

Discrepancies frequently occur among third-party labels from different annotators. Nevertheless, among these labels, one or a few may align with the SS. Therefore, samples where SS equals "Low" and "High" were extracted, and the consistency rate (rate1) of TS(Aas) with SS was compared for cases where SS = Low and SS = High. Additionally, samples were defined as consistent with SS if at least one third-party label

matched SS, and the consistency rate (rate2) was calculated. As shown in the table 5, rate2 is higher than rate1, particularly when SS = Low, where rate2 is significantly higher. This suggests that while most annotators may fail to accurately identify a participant's low emotion when the participant's emotion is genuinely low, a few annotators can still detect the participant's suppressed low emotion. However, this valuable information is not utilized when using TS(Aas) for prediction. To address this, the soft label method will be employed to learn from the labels provided by each annotator.

Table 5: Comparison of Consistency Rates between TS Hardlabel and TS Softlabel with SS(%)

	Hazumi1902		Hazumi1901	
SS	Low	High	Low	High
TS(Aas)	26	73	17	66
TS(soft)	52	89	44	85

In Chapter 3, the method for generating soft labels was explained. To ensure that the label with the highest probability in the soft label is consistent with TS(Aas) when comparing with the TS(Aas) model, an initial probability value of at least 0.5 is set at the position corresponding to the TS(Aas) label during the generation of the soft label. The specific implementation is illustrated in the Figure 7.

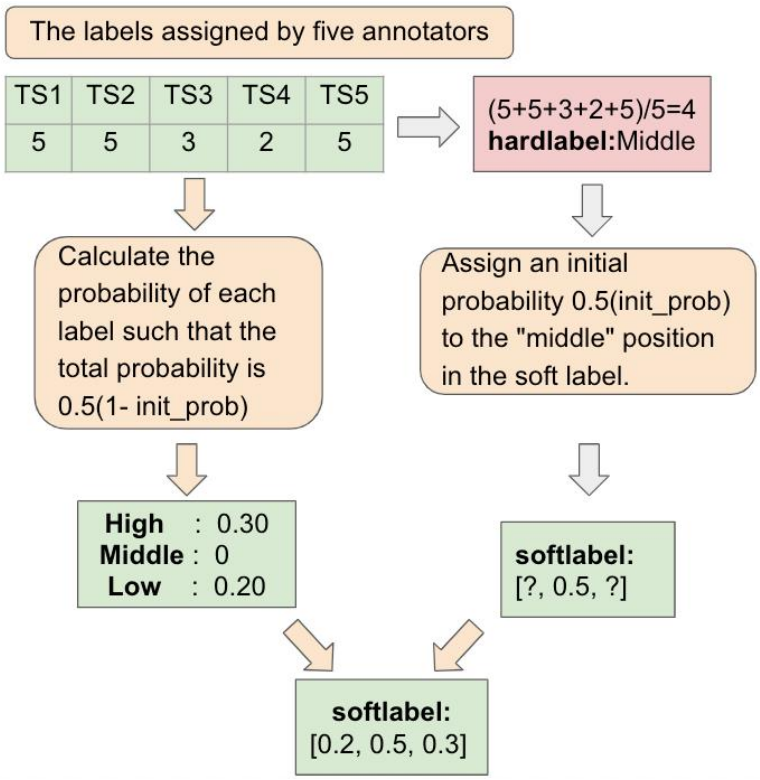

Figure 7: Example of Generate Softlabel with Initial Probability

The initial probabilities were set to [0.5, 0.6, 0.7, 0.8, 0.9] respectively, and the results were compared with the TS(As) model. The accuracy for all these initial probabilities was found to be higher than that of the TS(As) model, demonstrating the effectiveness of the soft label method. The results are presented in Table 6. Additionally, when the initial probability was set to 0, meaning the soft labels were generated purely based on the probability distribution without any initial probability setting, the performance, considering both accuracy and F1 score, was the highest.

Table 6: Experiments with TS Softlabel

	Hazumi1902		Hazumi1901	
	acc	f1	acc	f1
all with TS(Aas)	0.5250	0.3855	0.5239	0.3852
soft 0.5	0.5380	0.3945	0.5243	0.4031
soft 0.6	0.5379	0.3956	0.5247	0.4023
soft 0.7	0.5346	0.3935	0.5222	0.4027
soft 0.8	0.5329	0.3903	0.5265	0.4015
soft 0.9	0.5327	0.3964	0.5254	0.4049
soft 0	**0.5423**	0.3962	0.5253	**0.4068**

5.5 Experiments for Weighted Loss Dependency

Currently, it is evident that using TS soft labels to predict SS labels yields the best performance, surpassing the baseline. Previously, data was divided into two groups, Data(c) and Data(inc), based on the consistency of TS(As) with SS. By assigning higher loss weights to important samples, the model better learns the features of these samples, thereby improving overall accuracy. To explore the loss weight ratio between the two groups of samples, the loss weight of Data(c) was fixed at 1, and the loss weight of Data(inc) was adjusted to α. By adjusting α, the weight ratio between the two groups was varied, with α set to [0.25, 0.33, 0.5, 1, 2, 3, 4]. The experimental results are shown in Table 6. In both Hazumi1902 and Hazumi1911 datasets, the model performs best when α is set to 3. Additionally, there is a trend of improved model performance as α increases. However, continuously increasing α can lead to a deterioration in performance, as observed when α=4. The results are shown in Figure 8 and Figure 9.

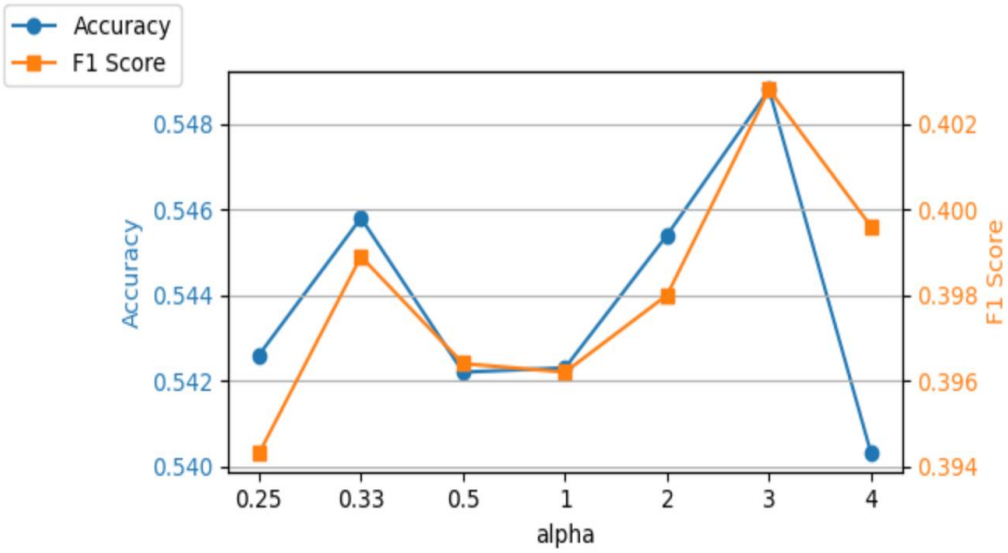

Figure 8: Experiments for Weighted Loss Dependency on Hazumi1902

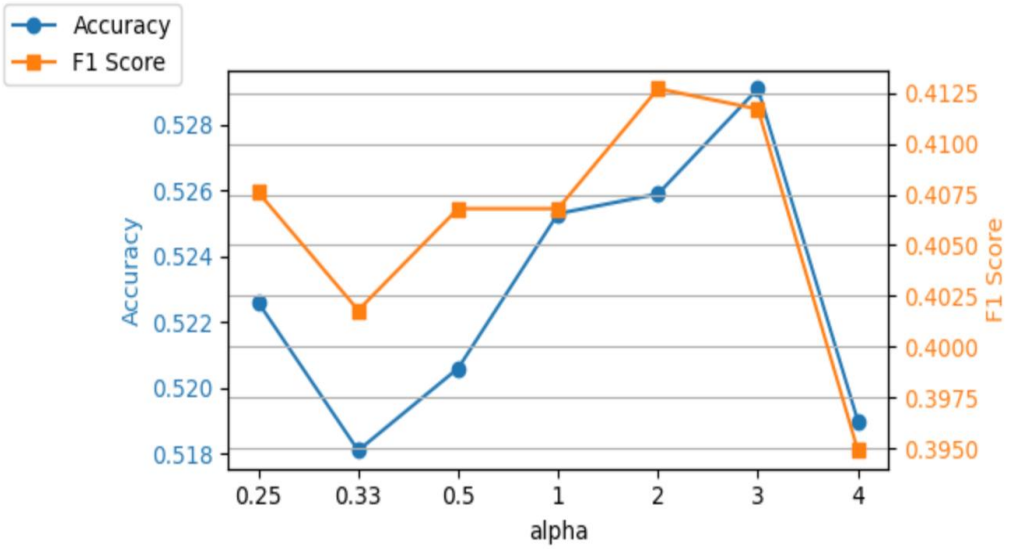

Figure 9: Experiments for Weighted Loss Dependency on Hazumi1911

5.6 Discussion

To summarize the results of the above experiments, the following 5 experimental results were selected for comparison. Additionally, a random guessing experiment was conducted for comparison with other experiments. Descriptions of each experiment are summarized in Table 7.

Table 7: Descriptions of Each Experiments

exp	plot_name	description
Random	rand	• Randomly label the test samples and compare with the ground truth SS
Baseline	base	• This experiment uses all the data without any modifications or special treatments. • Serves as the reference point for evaluating other methods.
Use TS(Aas)/SS consistent samples only	cons	• Data was divided based on whether TS(Aas) is consistent with SS. Only the consistent data was used.
Use TS softlabel and weighted loss with a=3	softlabel	• TS soft labels were used to predict SS labels. • Data was divided into two groups, Data(c) and Data(inc), based on TS(As) consistency with SS. • A weighted loss approach was applied where the loss weight for Data(c) was fixed at 1 and the loss weight for Data(inc) was set to 3.
Human level	human	• Represents the performance level of human annotators. • Serves as an upper bound or goal for model performance comparison.

Then plot the results using a bar chart, It can be seen that the softlabel method is the closest to human-level performance. The results are shown in Figure 10 and Figure 11.

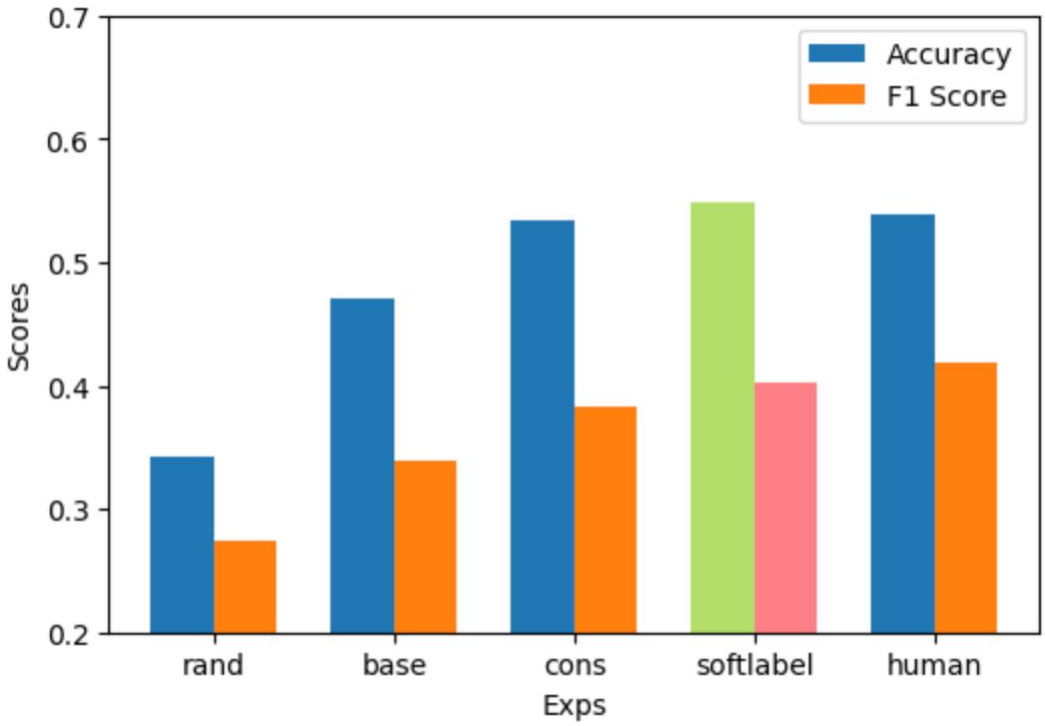

Figure 10: Comparisons of All Experiments on Hazumi1902

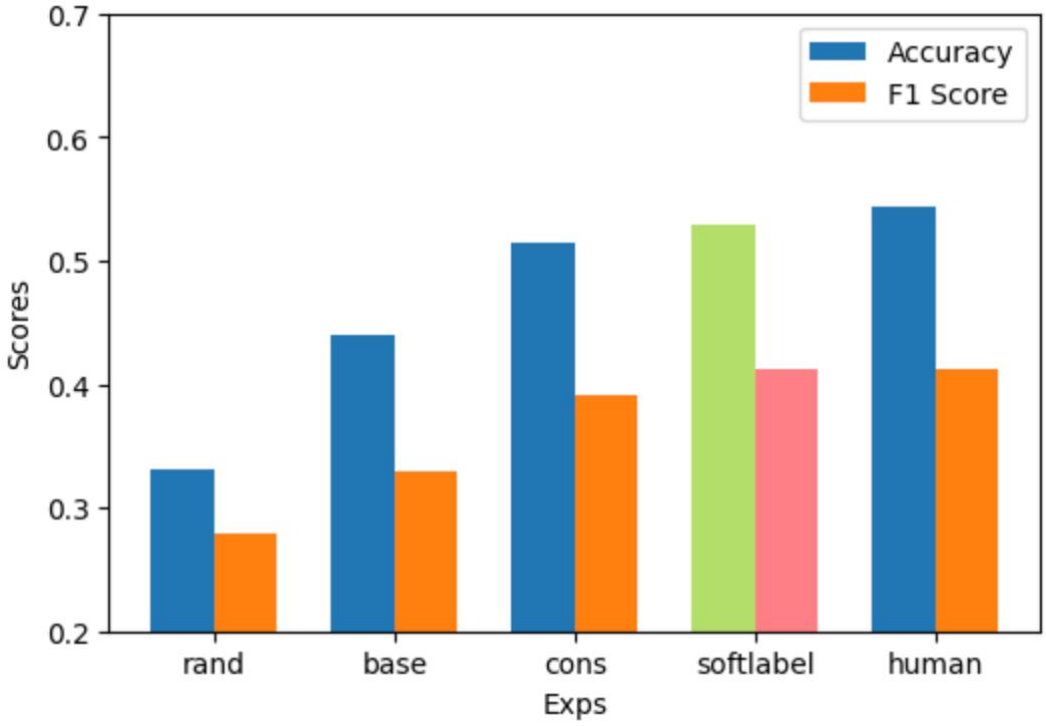

Figure 11: Comparisons of All Experiments on Hazumi1911

It can be observed that removing samples with inconsistent TS and SS labels significantly improves the model's performance. This suggests that when sample annotations conflict, the same input data may receive different labels, which can negatively impact the model's learning and accuracy. This inconsistency makes it difficult for the model to determine the correct classification boundary, causing confusion during the training process and reducing classification accuracy. In sentiment analysis tasks, inconsistent labels mean that different annotators have different interpretations and judgments of the sentiment for the same data point. This makes it challenging for the model to capture a consistent pattern of sentiment information, thereby affecting its classification performance.

When using TS instead of SS for training, the model essentially learns the pattern of human judgment regarding the participant's emotions. Since TS sample labels have higher consistency, it is understandable that the model's performance would be closer to human-level accuracy compared to the baseline. Additionally, when using TS soft labels for training, the accuracy on huzami1902 exceeds human levels, and the F1 score is relatively close to human levels. On hazumi1911, the F1 score surpasses human levels, and the accuracy is relatively close to human levels. This shows soft labels help to capture the differences between TS and SS annotations. When annotators disagree on a sentiment, a soft label can represent this ambiguity, reflecting the fact that the true sentiment might lie somewhere in between the various opinions.

Chapter 6
Conclusion

6.1 Summary

This thesis explored approaches to enhance self-sentiment (SS) estimation in multimodal dialogue systems, addressing persistent challenges in sentiment analysis. The focus was on employing soft labels to represent the complex and often ambiguous nature of human emotions. Soft labels provide a probabilistic representation of sentiment, capturing subtle variations in emotional states that hard labels might miss. Additionally, a weighted loss function strategy was proposed to address discrepancies between self-reported sentiments (SS) and third-party annotated sentiments (TS). By assigning different weights to samples based on the consistency between TS and SS labels, the approach aimed to improve the model's learning process and performance. Experiments were conducted using the Hazumi1902 and Hazumi1911 datasets, which provided diverse multimodal data for analysis. Results indicated that the proposed methods might offer improvements over traditional sentiment analysis approaches, especially in cases with significant sentiment label variation. The combination of the soft label method and weighted loss function demonstrated promising results, enhancing accuracy and reliability in sentiment estimation within multimodal dialogue contexts. These findings contribute to ongoing research in multimodal sentiment analysis by showcasing the potential of soft labels and weighted loss functions in capturing and modeling the complexities of human emotions. Future work could focus on refining the soft label generation process, integrating additional modalities such as physiological signals, and exploring the cross-cultural applicability of these methods.

6.2 Future Work

To advance our understanding and capabilities in sentiment analysis, future work can be approached from the following two perspectives:

More Effective Soft Label Generation Methods

Our research highlights the significant potential of soft label methods in capturing subtle emotional nuances, but there is considerable room for improvement. Future work could focus on developing algorithms that dynamically adjust soft labels based on context. This might involve incorporating factors such as conversational history, individual user characteristics, and environmental conditions to enhance the relevance and accuracy of the soft labels.

Sample Weighting Methods

While our weighted loss function strategy has demonstrated promising results, further optimization is needed. Future research could aim to create algorithms that automatically adjust sample weights during training. This adjustment would accommodate varying data distributions and model states, leading to more robust and adaptable models.

By exploring these areas, the goal is to develop more precise and reliable sentiment analysis systems. Such advancements could enhance model performance and offer new insights and applications in emotional intelligence within human-computer interactions.

References

[1] G. R. VandenBos, APA dictionary of psychology. American Psychological Association, 2007.

[2] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," Image and Vision Computing, vol. 65, pp. 3–14, 2017.

[3] J. D. Mayer and P. Salovey, "The intelligence of emotional intelligence," Intelligence, vol. 17, no. 4, pp. 433–442, 1993.

[4] J. Lin, W. Mao, and D. D. Zeng, "Personality-based refinement for sentiment classification in microblog," Knowledge-Based Systems, vol. 132, pp. 204–214, 2017.

[5] D. Goleman, Emotional intelligence. Bantam Books New York, 1995.

[6] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," Information Fusion, vol. 37, pp. 98–125, 2017.

[7] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in Proceedings of the 13th international conference on multimodal interfaces, 2011, pp. 169–176.

[8] V. P´erez-Rosas, R. Mihalcea, and L.-P. Morency, "Utterance-level multimodal sentiment analysis," in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2013, pp. 973–982.

[9] C. Clavel and Z. Callejas, "Sentiment analysis: from opinion mining to human-agent interaction," IEEE Transactions on affective computing, vol. 7, no. 1, pp. 74–93, 2015.

[10] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," IEEE Intelligent Systems, vol. 31, no. 6, pp. 82–88, 2016.

[11] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2236–2246.

[12] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," Language resources and evaluation, vol. 42, no. 4, pp. 335–359, 2008.

[13] K. P. Truong, D. A. Van Leeuwen, and F. M. De Jong, "Speechbased recognition of self-reported and observed emotion in a dimensional space," Speech communication, vol. 54, no. 9, pp. 1049–1063, 2012.

[14] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. IEEE Trans. on Pattern Analysis and Machine Intelligence 41, 2 (Feb 2019), 423–443.

[15] Dan Bohus and Eric Horvitz. 2009. Learning to Predict Engagement with a Spoken Dialog System in Open-world Settings. In Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL). 244–252.

[16] Yukiko I. Nakano and Ryo Ishii. 2010. Estimating User's Engagement from

Eye-gaze Behaviors in Human-agent Conversations. In Proc. ACM International Conference on Intelligent User Interfaces (IUI). 139–148.
[17] Mohammed Ehsan Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W Picard. 2013. Mach: My automated conversation coach. In Proc. ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp). ACM, 697–706.
[18] Hiroki Tanaka, Hideki Negoro, Hidemi Iwasaka, and Satoshi Nakamura. 2017. Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders. PloS one 12, 8 (2017), e0182151.
[19] Yuya Chiba, Masashi Ito, Takashi Nose, and Akinori Ito. 2014. User Modeling by Using Bag-of-Behaviors for Building a Dialog System Sensitive to the Interlocutor's Internal State. In Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL). 74–78.
[20] Takatsugu Hirayama, Yasuyuki Sumi, Tatsuya Kawahara, and Takashi Matsuyama. 2011. Info-concierge: Proactive multi-modal interaction through mind probing. In The Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2011).
[21] Klaus Weber, Hannes Ritschel, Ilhan Aslan, Florian Lingenfelser, and Elisabeth André. 2018. How to Shape the Humor of a Robot - Social Behavior Adaptation Based on Reinforcement Learning. In Proc. ACM International Conference on Multimodal Interaction (ICMI). 154–162.
[22] Setareh Nasihati Gilani, David Traum, Arcangelo Merla, Eugenia Hee, Zoey Walker, Barbara Manini, Grady Gallagher, and Laura-Ann Petitto. 2018. Multimodal Dialogue Management for Multiparty Interaction with Infants. In Proc. ACM International Conference on Multimodal Interaction (ICMI). 5–13.
[23] Joan-Isaac Biel, Vagia Tsiminaki, John Dines, and Daniel Gatica-Perez. 2013. Hi YouTube! Personality Impressions and Verbal Content in Social Video. In Proc. ACM International Conference on Multimodal Interaction (ICMI). 119–126.
[24] Maarten Brilman and Stefan Scherer. 2015. A Multimodal Predictive Model of Successful Debaters or How I Learned to Sway Votes. In Proc. ACM International Conference on Multimedia. 149–158.
[25] Hongliang Yu, Liangke Gui, Michael Madaio, Amy Ogan, Justine Cassell, and Louis-Philippe Morency. 2017. Temporally Selective Attention Model for Social and Affective State Recognition in Multimedia Content. In Proc. ACM International Conference on Multimedia. 1743–1751.
[26] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018. Multi-attention Recurrent Network for Human Communication Comprehension. In Proc. Association for the Advancement of Artificial Intelligence (AAAI).
[27] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory Fusion Network for Multi-view Sequential Learning. In Proc. Association for the Advancement of Artificial Intelligence (AAAI).
[28] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP). 1103–1114.

[29] Xavier Alameda-Pineda, Yan Yan, Elisa Ricci, Oswald Lanz, and Nicu Sebe. 2015. Analyzing Free-Standing Conversational Groups: A Multimodal Approach. In Proc. ACM International Conference on Multimedia. 5–14.
[30] Ekin Gedik and Hayley Hung. 2018. Detecting Conversing Groups Using Social Dynamics from Wearable Acceleration: Group Size Awareness. Proc. ACM Interact. Mob. Wearable Ubiquitous Technology 2, 4, Article 163 (2018), 24 pages.
[31] Yuki Hirano, Shogo Okada, Haruto Nishimoto, and Kazunori Komatani. 2019. Multitask Prediction of Exchange-Level Annotations for Multimodal Dialogue Systems. In Proc. ACM International Conference on Multimodal Interaction (ICMI). 85–94.
[32] Hirano, S. Okada, and K. Komatani, "Recognizing social signals with weakly supervised multitask learning for multimodal dialogue systems," in Proc. Int. Conf. Multimodal Interaction, 2021, pp. 141–149.
[33] Koji Inoue, Divesh Lala, Katsuya Takanashi, and Tatsuya Kawahara. 2018. Latent character model for engagement recognition based on multimodal behaviors. In Proc. International Workshop on Spoken Dialogue Systems (IWSDS).
[34] Shiro Kumano, Kazuhiro Otsuka, Dan Mikami, Masafumi Matsuda, and Junji Yamato. 2015. Analyzing Interpersonal Empathy via Collective Impressions. IEEE Trans. on Affective Computing 6, 4 (2015), 324–336.
[35] Derya Ozkan and Louis-Philippe Morency. 2013. Latent Mixture of Discriminative Experts. IEEE Trans. on Multimedia 15, 2 (2013), 326–338.
[36] Qianli Xu, Liyuan Li, and Gang Wang. 2013. Designing Engagement-Aware Agents for Multiparty Conversations. In Proc. ACM CHI Conference (CHI '13). 2233–2242.
[37] Reza Lotfian and Carlos Busso. 2019. Curriculum Learning for Speech Emotion Recognition From Crowdsourced Labels. IEEE/ACM Trans. Audio, Speech and Lang. Proc. 27, 4 (2019), 815–826.
[38] Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 175–190, Seattle, United States. Association for Computational Linguistics.
[39] Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future, pages 15–21, Online. Association for Computational Linguistics.
[40] Emily Jamison and Iryna Gurevych. 2015. Noise or additional information? leveraging crowdsource annotation item agreement for natural language tasks. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 291–297, Lisbon, Portugal. Association for Computational Linguistics.
[41] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. Transactions of the Association for Computational Linguistics, 10:92–110.
[42] Ben Wu, Yue Li, Yida Mu, Carolina Scarton, Kalina Bontcheva, and Xingyi Song.

2023. Don’t waste a single annotation: Improving single-label classifiers through soft labels. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 5347–5355, Singapore. Association for Computational Linguistics.
[43] Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Crowdsourcing semantic label propagation in relation classification. In Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), pages 16–21, Brussels, Belgium. Association for Computational Linguistics.
[44] Kazunori Komatani and Shogo Okada. 2021. Multimodal Human-Agent Dialogue Corpus with Annotations at Utterance and Dialogue Levels. In Proc. International Conference on Affective Computing and Intelligent Interaction (ACII).
[45] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre- training of deep bidirectional transformers for language under- standing,” in NAACL-HLT (1), 2019.
[46] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Im- proving language understanding with unsupervised learning,” Technical report, OpenAI, 2018.
[47] T. Akahori, K. Dohsaka, M. Ishii, and H. Ito, “Efficient creation of japanese tweet emotion dataset using sentence-final expressions,” in 2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech). IEEE, 2021, pp. 501–505.
[48] B. Schuller, S. Steidl, and A. Batliner, “The interspeech 2009 emotion challenge,” in Tenth Annual Conference of the International Speech Communication Association, 2009.
[49] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, “OpenFace 2.0: Facial behavior analysis toolkit,” in 13th IEEE International Conference on Automatic Face & Gesture Recognition. IEEE, 2018, pp. 59–66.
[50] E.FriesenandP.Ekman,“Facialactioncodingsystem:atechnique for the measurement of facial movement,” Palo Alto, vol. 3, 1978.