Forecasting Implicit Emotions Elicited in Conversations

Yurie Koga **Shunsuke Kando** **Yusuke Miyao**
Department of Computer Science
The University of Tokyo
{ykrasp7isweet, skando, yusuke}@is.s.u-tokyo.ac.jp

Abstract

This paper aims to forecast the implicit emotion elicited in the dialogue partner by a textual input utterance. Forecasting the interlocutor's emotion is beneficial for natural language generation in dialogue systems to avoid generating utterances that make the users uncomfortable. Previous studies forecast the emotion conveyed in the interlocutor's response, assuming it will explicitly reflect their elicited emotion. However, true emotions are not always expressed verbally. We propose a new task to directly forecast the implicit emotion elicited by an input utterance, which does not rely on this assumption. We compare this task with related ones to investigate the impact of dialogue history and one's own utterance on predicting explicit and implicit emotions. Our result highlights the importance of dialogue history for predicting implicit emotions. It also reveals that, unlike explicit emotions, implicit emotions show limited improvement in predictive performance with one's own utterance, and that they are more difficult to predict than explicit emotions. We find that even a large language model (LLM) struggles to forecast implicit emotions accurately.

1 Introduction

Dialogue system is a key application of natural language generation. For dialogue systems, forecasting user reactions to generated utterances is beneficial for preventing potentially offensive responses. In this research, we introduce the task of forecasting the implicit emotion elicited in the dialogue partner by a textual input utterance.

Several previous studies (Hasegawa et al., 2013; Li et al., 2020, 2021a; Zhang et al., 2021) forecast the emotion of a dialogue partner by using speaker emotion datasets. The emotion labels in these datasets represent the emotions expressed in utterances, which means they assumed the emotion elicited in the interlocutor will explicitly be conveyed in their response. However, this does not always hold true, as individuals may hide their true emotions. Another study (Shen et al., 2020) directly predicted elicited implicit emotions using both the preceding and subsequent context, but the latter is usually unavailable in dialogue systems.

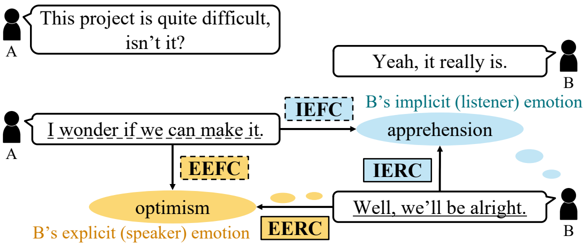
A
This project is quite difficult, isn't it?
Yeah, it really is.
B
IEFC
B's implicit (listener) emotion
A
I wonder if we can make it.
apprehension
EEFC
IERC
optimism
Well, we'll be alright.
B
B's explicit (speaker) emotion
EERC

Figure 1: Example of the four emotion classification tasks we discuss. The emotions are taken from Plutchik's wheel of emotions (Plutchik, 2001). In this conversation, while B feels apprehension because of A's anxious utterance, "I wonder if we can make it.", B expresses optimism in his utterance to encourage himself.

Table 1: The classification of the four tasks.

	Recognition	Forecasting
Explicit	EERC	EEFC
Implicit	IERC	IEFC

We propose a new forecasting task, which uses a listener emotion dataset and only the preceding dialogue history. We compare this task with three related tasks by fine-tuning DistilRoBERTa (Liu et al., 2019; Sanh et al., 2020) for each one. This comparison explores the impact of dialogue history and one's own utterance on the difficulty of predicting explicit and implicit emotions. The four tasks are defined by two criteria (explicit/implicit, recognition/forecasting) as described in Figure 1 and Table 1. In the following, the term prediction is used to refer to both recognition and forecasting. Explicit tasks predict speaker emotions expressed in utterances, while implicit ones predict listener emotions, which are not always expressed. Recognition

Proceedings of the 17th International Natural Language Generation Conference, pages 145–152
September 23–27, 2024. ©2024 Association for Computational Linguistics

tasks predict emotions from one's own utterance, whereas forecasting ones predict emotions from the preceding utterance of the dialogue partner. The main task we mentioned above corresponds to an implicit and forecasting one. We experiment with three settings for each task, varying the amount of dialogue history to feed the model. In addition, we fine-tune Llama 2 (Touvron et al., 2023) for the main task (implicit & forecasting) to examine whether a large language model (LLM) can perform this task.

Analysis of our results suggests three implications: (1) the importance of dialogue history in predicting implicit emotions, (2) the limited improvement in the predictive performance of implicit emotions with one's own utterance compared to explicit ones, (3) the greater difficulty of predicting implicit emotions over explicit ones. We also observed that forecasting implicit emotions is challenging even for an LLM.

2 Related Work

Some previous studies have attempted the forecasting task, which is to predict the dialogue partner's emotion. They incorporated commonsense knowledge (Li et al., 2021b; Fujimoto and Ito, 2023) or emotional persistence and contagiousness (Li et al., 2020, 2021a) in addition to dialogue history (Hasegawa et al., 2013). Their task differs from ours as they employed speaker emotion datasets for training and evaluation.

Listener emotion datasets are used by two studies. The first one (Shen et al., 2020), which created the MEmoR dataset, predicted both the speakers' explicit emotions and the listeners' implicit emotions based on multimodal and personality information. The results suggest that predicting listeners' emotions is more difficult than predicting speakers' emotions. This work differs from ours as it used the subsequent context, which is unavailable in dialogue systems.

The other study (Gong et al., 2023), which created the reconstructed MEmoR dataset, built a positive emotion elicitation dialogue system. MEmoR was reconstructed so that all the emotion labels could be inferred from the textual information alone. The dataset was used to train a latent variable to control the emotional tone of utterances. Instead, we train a model to forecast implicit emotions directly. Implementing such a model in dialogue systems will enhance their interpretability.

3 Emotion Classification Tasks

We focus on the task of forecasting the implicit emotion elicited by an utterance in its listener and compare it to three related tasks. The four tasks are divided into explicit and implicit emotion predictions, and further into recognition and forecasting. Here, the *speaker emotion* refers to the emotion explicitly expressed in an utterance, and the *listener emotion* refers to the implicit emotion elicited by an utterance. Figure 1 and Table 1 show an overview.

3.1 Explicit Emotion Prediction

Explicit emotions refer to those explicitly conveyed in the utterances. The prediction targets are the speaker emotion labels (e.g., "optimism" in Figure 1), as those are inferred from the utterances and thus can be considered as expressed in them.

Explicit Emotion Recognition in Conversations (EERC) EERC predicts the speaker emotion from the speaker's corresponding utterance (e.g., B's speaker emotion "optimism" from B's utterance "Well, we'll be alright." in Figure 1). In addition to the utterance itself, dialogue history and speaker information are often considered (Ghosal et al., 2019; Poria et al., 2019b). We utilize only dialogue history in our experiments to make them simple.

Explicit Emotion Forecasting in Conversations (EEFC) EEFC predicts the speaker emotion of the next utterance from the current utterance (e.g., B's next speaker emotion "optimism" from A's current utterance "I wonder if we can make it." in Figure 1). Unlike EERC, the target utterance to predict the emotion is yet to come. Dialogue history is often used as a clue (Hasegawa et al., 2013; Li et al., 2020, 2021a,b; Fujimoto and Ito, 2023), and we use it in our experiments.

3.2 Implicit Emotion Prediction

Implicit emotions refer to true emotions, which are not necessarily expressed in the utterances. The prediction targets are the listener emotion labels (e.g., "apprehension" in Figure 1). To the best of our knowledge, predicting these emotions from the preceding context alone has not been studied yet.

Implicit Emotion Recognition in Conversations (IERC) IERC predicts the current listener emotion from the listener's next utterance (e.g., B's elicited listener emotion "apprehension" from B's next utterance "Well, we'll be alright." in Figure 1).

Table 2: The way emotions are labeled in conversations. u_i^X is X's utterance in the i-th turn. e_s^X is X's speaker emotion expressed in the utterance in the same line and e_l^X is the emotion elicited in listener X by the utterance in the same line. In Figure 1, u_n^A corresponds to "I wonder if we can make it.", u_n^B to "Well, we'll be alright.", e_s^B to "optimism", and e_l^B to "apprehension". Speaker and listener emotions are annotated in DailyDialog and reconstructed MEmoR, respectively.

Utterance	Speaker Emotion	Listener Emotion
u_1^A	-	-
u_1^B	-	-
$\dots$	$\dots$	$\dots$
u_n^A	-	e_l^B
u_n^B	e_s^B	-

Table 3: Task definitions. We used the space character for concatenation, represented here as ":".

Task	Input			Output
	full history	last uttr	no history	
EERC	$u_1^A : u_1^B : \dots : u_n^B$	$u_n^A : u_n^B$	u_n^B	e_s^B
EEFC	$u_1^A : u_1^B : \dots : u_n^A$	$u_{n-1}^B : u_n^A$	u_n^A	e_s^B
IERC	$u_1^A : u_1^B : \dots : u_n^B$	$u_n^A : u_n^B$	u_n^B	e_l^B
IEFC	$u_1^A : u_1^B : \dots : u_n^A$	$u_{n-1}^B : u_n^A$	u_n^A	e_l^B

Implicit Emotion Forecasting in Conversations (IEFC) IEFC predicts the implicit emotion elicited in the listener by an input utterance (e.g., B's elicited listener emotion "apprehension" from A's utterance "I wonder if we can make it." in Figure 1). This task is our primary focus. It is sometimes approximated by EEFC (Hasegawa et al., 2013; Li et al., 2020, 2021a,b; Fujimoto and Ito, 2023), a task to predict the next speaker emotion (e.g., "optimism" in Figure 1) from the same input. These two are the same if the emotion elicited in the listener is always expressed in the listener's next utterance, but humans sometimes hide their emotions. For example, in Figure 1, B's listener emotion "apprehension" differs from B's next speaker emotion "optimism".

4 Experiment

4.1 Task Definition

We experimented with four tasks: EERC, EEFC, IERC, and IEFC, mainly focusing on IEFC. Table 2 and 3 show the emotion labeling and the task definitions, respectively. For each task, we experimented with three different input variations: full history, last utterance, and no history, varying the amount of dialogue history to concatenate.

4.2 Dataset

We used two different datasets for the explicit and implicit tasks because no dataset has both speaker and listener emotion annotations based solely on textual information.

For explicit tasks (EERC, EEFC), we used DailyDialog (Li et al., 2017), which consists of daily life dyadic textual conversations. The utterances are annotated with seven emotion labels: Ekman's six primary emotions (anger, disgust, fear, happiness, sadness, surprise) (Ekman, 1992) and no emotion.

For implicit tasks (IERC, IEFC), we employed reconstructed MEmoR (Gong et al., 2023). It is extracted from MEmoR (Shen et al., 2020), a multimodal dataset of dialogues from the TV Show "The Big Bang Theory". In MEmoR, both the speaker and listener emotion labels are annotated to each utterance using multimodal information. During reconstruction (Gong et al., 2023), all the non-textual information and speaker emotion labels were removed, and the listener emotion labels were ensured to be inferred solely from the text dialogue history. The emotion labels are positive, negative, and neutral.

4.3 Data Preprocessing

We performed two data preprocessings: two-party conversation filtering and label conversion.

First, we extracted two-party conversations from reconstructed MEmoR, as we focus on two-party situations. We used DailyDialog as it is.

Then, we converted the emotion labels of DailyDialog to positive, negative, or neutral to match the categories of reconstructed MEmoR. Happiness was mapped to positive, no emotion to neutral, and anger, disgust, fear, and sadness were mapped to negative. Surprise was excluded from prediction targets because it can indicate either positive or negative emotions in Ekman's six primary emotions (Poria et al., 2019a). Note that the labels are biased toward neutral, with 84.6% of labels in DailyDialog and 80.3% in reconstructed MEmoR being neutral.

See Appendix A for more detail on the preprocessed datasets.

4.4 Training

We fine-tuned DistilRoBERTa-base[1] (Liu et al., 2019; Sanh et al., 2020) for each emotion clas-

[1] https://huggingface.co/distilbert/distilroberta-base

sification task. To further explore the performance of an LLM on IEFC, we fine-tuned Llama-2-13b-hf[2] (Touvron et al., 2023) for IEFC. For the DistilRoBERTa model, we experimented under two settings: using all available train data for each task, and standardizing the train data size to 4,767 (the minimum train data size among all the tasks; see Table 6) across all the tasks. See Appendix B for the hyperparameters.

Due to the biased label distribution towards neutral in both datasets, we trained with a weighted loss in every experiment. The detailed formula is:

$$\text{WeightedCrossEntropyLoss}(\boldsymbol{p}, \boldsymbol{y}) = -\sum_{i=1}^{n} \frac{\sum_{j=1}^{n} C_j}{C_i} y_i \log p_i,$$

where $\boldsymbol{p}$ is the predicted probabilities of the classes, $\boldsymbol{y}$ is the correct one-hot vector, n is the number of classes, and C_i is the number of data in class i.

Evaluation Metrics We evaluated the models using macro-F1 score and F1 w/o neutral score, which is the average of the F1 scores of the positive and negative classes. We employed them to assess the models' ability to predict the positive and negative labels in datasets with a bias toward neutral.

5 Results

Figure 2 displays the macro-F1 and F1 w/o neutral scores of DistilRoBERTa across the four tasks. The left figures show the results using all available training data for each task, while the right ones show the results using a standardized 4,767 training samples for all the tasks. Each score point and its corresponding error bar represent the average and standard error of five trials with different random seeds for train data selection.

Overall, the results with dialogue history outperform those without it, especially for implicit tasks. This indicates that the context is important in predicting implicit emotions. As for IEFC with 4,767 training samples, the last utterance setting yielded better results than the full-history setting. This might be because the elicited implicit emotion is greatly influenced by the person's last utterance (e.g., B's utterance "Yeah, it really is." in Figure 1), and can be confused by earlier dialogue history (e.g., "This project is quite difficult, isn't it?" in Figure 1).

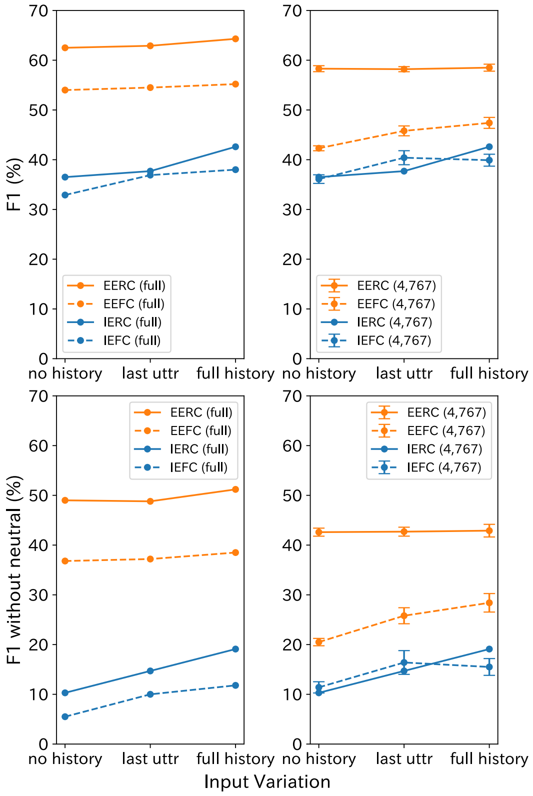

Figure 2: Macro-F1 (above) and F1 w/o neutral (below) scores of each task. The random baseline of the macro-F1 score is 24.6% for EERC and EEFC, and 22.8% for IERC and IEFC. The random baseline of the F1 w/o neutral score is 13.2% for EERC and EEFC, and 10.0% for IERC and IEFC.

5.1 Recognition vs. Forecasting

As for the explicit tasks, the EERC results significantly outperform those of EEFC. This may be because the speaker's explicit emotion is easier to predict from their own utterance than from the dialogue partner's utterance. Conversely, as for the implicit tasks, the IERC results are only marginally better than those of IEFC, even when feeding the entire dialogue history to the model.

5.2 Explicit vs. Implicit

The results of EERC and EEFC surpass those of IERC and IEFC, respectively. When the emotion elicited in the listener is expressed in their next utterance, there is no difference between EERC and IERC, or EEFC and IEFC. Given this, the result suggests that the listener's emotion is not always reflected in the subsequent utterance, making implicit emotion prediction more challenging than explicit emotion prediction. Additionally, it indicates

[2] https://huggingface.co/meta-llama/Llama-2-13b-hf

Table 4: The F1 w/o neutral scores of Llama 2 for IEFC. The random baseline is 10.0%.

Input Variation	F1 w/o neutral score
no history	12.4%
last utterance	22.5%
full history	**27.7%**

that IEFC, the task that we proposed, which has a more realistic setting, is actually more difficult than EEFC, the focus of previous studies. Note that this comparison might be limited as the datasets for the explicit and implicit tasks differ in this experiment.

5.3 LLM Results

Table 4 shows the F1 w/o neutral scores of Llama 2 for IEFC using all available train data. Although Llama 2 performs better than DistilRoBERTa, it still struggles with forecasting implicit emotions.

6 Conclusion

We proposed a new task to forecast the implicit emotion elicited in the listener by an input utterance, and analyzed its difficulty by comparing it with three related tasks. The analysis suggests three points: (1) dialogue history is important for predicting implicit emotions, (2) unlike explicit emotions, implicit emotions show limited improvement in predictive performance with one's own utterance, (3) implicit emotions are more challenging to predict than explicit ones. Additionally, we fine-tuned Llama 2 for the new task and found it struggles to accurately forecast elicited implicit emotions.

As future work to improve its performance, possible directions include applying prompt engineering techniques or using other large language models. Incorporating personality information (Shen et al., 2020) or commonsense knowledge (Li et al., 2021b; Fujimoto and Ito, 2023) is also a promising approach. Personalities will be particularly important for this task, since the emotion elicited in the listener by an utterance is likely to vary with the personality of the listener (Shen et al., 2020). Further, this task can be extended to multi-party conversations and situations with multimodal information.

Supplementary Materials Availability Statement: We will make the source code available at GitHub[3]. DailyDialog is available at HuggingFace[4]. Reconstructed MEmoR (Gong et al., 2023) is not openly published due to the license of the original MEmoR dataset.

Acknowledgments

We would like to express our gratitude to Qiang Zhang for his support. This work was supported by the Institute of AI and Beyond of the University of Tokyo.

References

Paul Ekman. 1992. An argument for basic emotions. Cognition and Emotion, 6(3-4):169–200.

Takumi Fujimoto and Takayuki Ito. 2023. Emotion prediction based on conversational context and commonsense knowledge graphs. In Advances and Trends in Artificial Intelligence. Theory and Applications, pages 407–412, Cham. Springer Nature Switzerland.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 154–164, Hong Kong, China. Association for Computational Linguistics.

Ziwei Gong, Qingkai Min, and Yue Zhang. 2023. Eliciting rich positive emotions in dialogue generation. In Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023), pages 1–8, Toronto, Canada. Association for Computational Linguistics.

Takayuki Hasegawa, Nobuhiro Kaji, Naoki Yoshinaga, and Masashi Toyoda. 2013. Predicting and eliciting addressee's emotion in online dialogue. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 964–972, Sofia, Bulgaria. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. Preprint, arXiv:2106.09685.

Dayu Li, Yang Li, and Suge Wang. 2020. Interactive double states emotion cell model for textual dialogue emotion prediction. Knowledge-Based Systems, 189:105084.

[3] https://github.com/mynlp/Forecasting_Implicit_Emotions

[4] https://huggingface.co/datasets/daily_dialog

Dayu Li, Xiaodan Zhu, Yang Li, Suge Wang, Deyu Li, Jian Liao, and Jianxing Zheng. 2021a. Emotion inference in multi-turn conversations with addressee-aware module and ensemble strategy. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3935–3941, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dayu Li, Xiaodan Zhu, Yang Li, Suge Wang, Deyu Li, Jian Liao, and Jianxing Zheng. 2021b. Enhancing emotion inference in conversations with commonsense knowledge. Knowledge-Based Systems, 232:107449.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. Preprint, arXiv:1907.11692.

Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. American Scientist, 89(4):344–350.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019a. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019b. Emotion recognition in conversation: Research challenges, datasets, and recent advances. IEEE Access, 7:100943–100953.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. Preprint, arXiv:1910.01108.

Guangyao Shen, Xin Wang, Xuguang Duan, Hongzhi Li, and Wenwu Zhu. 2020. Memor: A dataset for multimodal emotion reasoning in videos. In Proceedings of the 28th ACM International Conference on Multimedia, MM '20, page 493–502, New York, NY, USA. Association for Computing Machinery.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. Preprint, arXiv:2307.09288.

Rui Zhang, Zhenyu Wang, Zhenhua Huang, Li Li, and Mengdan Zheng. 2021. Predicting emotion reactions for human–computer conversation: A variational approach. IEEE Transactions on Human-Machine Systems, 51(4):279–287.

Table 5: Label distribution of datasets. The labels in both datasets are biased towards neutral.

Dataset	DailyDialog	reconstructed MEmoR
Positive	12.7%	8.7%
Neutral	84.6%	80.3%
Negative	2.7%	11.0%
Total	100.0%	100.0%

Table 6: Train/Valid/Test split.

Dataset	**Task**	**Train**	**Valid**	**Test**
DailyDialog	EERC	85,570	7,962	6,632
	EEFC	74,548	6,973	6,632
reconstructed	IERC	4,767	585	573
MEmoR	IEFC	7,810	742	573

A Dataset Details

We show the label distribution of each dataset in Table 5 and the number of data for each task in Table 6. The datasets were split in the same way as the original data for both DailyDialog and reconstructed MEmoR. The train and validation data sizes for EEFC are smaller than those for EERC, and IERC than IEFC. This is because EEFC and IERC require two annotated utterances as the input (i.e., the current utterance and the next emotion, the current emotion and the next utterance). As for the test data, we used the same data for EERC and EEFC, and for IERC and IEFC to compare the results between these tasks.

B Hyperparameters

The hyperparameters are shown in Table 7. All the models were trained with one GPU (NVIDIA A100). At the end of the training of each task, we loaded the model of the epoch that achieved the highest macro-F1 score on the validation dataset. We fine-tuned Llama 2 using LoRA (Hu et al., 2021).

Table 7: Hyperparameters.

Model	**Task**	**Input Variation**	**Learning Rate**	**Batch Size**	**Epoch**
Llama-2-13b-hf	IEFC	full history	1e-5	4	10
		last uttr	2e-5	2	
		no history	2e-5	1	
DistilRoBERTa-base	EERC	all	warmup from 0 to 5e-05	64	40
	EEFC			64	60
	IERC			128	40
	IEFC			128	40