Mitigating Linguistic Artifacts in Emotion Recognition for Conversations from TV Scripts to Daily Conversations

Donovan Ong[*,1], Shuo Sun[2], Jian Su[2], Bin Chen[2]
[1]Nanyang Technological University
[2]Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
S210063@e.ntu.edu.sg, {Sun_Shuo,sujian,bchen}@i2r.a-star.edu.sg

Abstract

Emotion Recognition in Conversations (ERC) is a well-studied task with numerous potential real-world applications. However, existing ERC models trained on the MELD dataset derived from TV series struggle when applied to daily conversation datasets. A closer examination of the datasets unveils the prevalence of linguistic artifacts, such as repetitions and interjections in TV scripts, which ERC models may exploit when making predictions. To address this issue, we explore two techniques aimed at reducing the reliance of ERC models on these artifacts: 1) using contrastive learning to prioritize emotional features over dataset-specific linguistic style and 2) refining emotion predictions with pseudo-emotion intensity score. Our experiment results show that reducing reliance on the linguistic style found in TV transcripts could enhance the model's robustness and accuracy in diverse conversational contexts.

Keywords: Emotion Recognition for Conversations, Adaptability

1. Introduction

Given a dialogue with multiple utterances, the goal of Emotion Recognition in Conversations (ERC) is to identify the emotion expressed by the speaker for each utterance. ERC models analyze the words and phrases used in each utterance to extract valuable clues about the speaker's emotional state. Recent studies have also demonstrated the significant benefits of incorporating conversation context and commonsense knowledge (Ghosal et al., 2019; Lee and Lee, 2022). ERC has many potential applications, including improving customer service, enhancing personal relationships, and diagnosing and treating mental health conditions.

Since its inception as a research topic, several datasets (Zahiri and Choi, 2018; Hsu et al., 2018; Poria et al., 2019) have been introduced, covering a range of conversation settings for ERC in English. However, existing works only focus on training and testing models on the same datasets, and there is no prior work on adaptability (Ghosal et al., 2019; Zhong et al., 2019; Ghosal et al., 2020; Lee and Lee, 2022). Research in the adaptability of ERC has been hindered by the challenges of unifying datasets with different emotion taxonomies and conversation settings, including TV series (Hsu et al., 2018; Poria et al., 2019), daily conversations (Li et al., 2017), and social media (Chatterjee et al., 2019).

This paper aims to address this knowledge gap by presenting a preliminary investigation into the adaptability of ERC models. We observe a significant disparity in performance when applying a model trained on TV transcripts to daily conversation data when using the same set of emotion labels. Upon closer examination of the datasets, we found evidence of linguistics artifacts that the models exploit to make predictions. To mitigate this issue, we delve into techniques such as contrastive learning and emotional intensity calibration, effectively reducing the models' reliance on these artifacts. Additionally, our findings demonstrate that these techniques are applicable across diverse conversational topics.

2. Methodology

Label	MELD	DailyDialog
Neutral	47.0%	83.1%
Joy	16.8%	12.5%
Surprise	11.9%	1.8%
Anger	11.7%	1.0%
Sadness	7.3%	1.1%
Disgust	2.6%	0.3%
Fear	2.6%	0.2%

Table 1: Label distribution of MELD and DailyDialog.

Dataset We have selected MELD (Poria et al., 2019) and DailyDialog (Li et al., 2017) as our evaluation datasets since they employ the same set of emotion labels (joy, anger, sadness, fear, disgust, surprise, and neutral). This commonality enables us to make a direct comparison of the model performance across different types of conversational data. MELD and DailyDialog contain 9,989/1,109/2,610

[*]Work done when Donovan was working at I^{2}R

LREC-COLING 2024, pages 11319–11324
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

and 87,167/8,069/7,740 train/dev/test utterances, respectively. The distribution of emotion labels in these datasets is presented in Table 1.

Evaluation Metric Previous studies have utilized various evaluation metrics for these datasets. Specifically, the official evaluation metric for the MELD is the *weighted-F1*, which calculates the mean of per-label F1 scores normalized by the number of samples in each label. On the other hand, DailyDialog employs *micro-F1*, which computes the global F1 score over all labels, excluding the "neutral" category. In this paper, we use the macro-F1 metric, which assigns equal weight to each class. This approach ensures a fair comparison across datasets with varying label distributions, allowing us to gauge model performance more effectively.

Baseline Our baseline model is an enhanced version of the c-LSTM model (Poria et al., 2017) where we replace the LSTM-based utterance encoder with RoBERTa-large (Liu et al., 2019), a popular pre-trained language model that has been proven to be effective in recent works on ERC. The model comprises a hierarchical structure that encodes each utterance into a single vector representation with RoBERTa, followed by an RNN-based dialogue encoder that considers the sequential relationship between utterances. A linear layer (CLS) then classifies the output dialogue representation into one of the emotion categories. We also replaced the LSTM with GRU for the dialogue encoding component.

We consider the hierarchical model over recent state-of-the-art models for its simplicity and its independence from external components, such as commonsense knowledge as seen in KET (Zhong et al., 2019) and COSMIC (Ghosal et al., 2020), or speaker identity in CoMPM (Lee and Lee, 2022). Moreover, the hierarchical model exhibits competitive performance (65.3 weighted-F1) against the state-of-the-art model, CoMPM (Lee and Lee, 2022) (66.5 weighted-F1) on MELD despite not being optimized on weighted-F1.

3. Adaptability Study

3.1. Performance

As seen in Table 2, evaluation of the MELD-trained model on the DailyDialog test set revealed a 14.9% absolute difference in macro-F1 compared to its in-distribution result. We also observed a significant performance gap when trained with equal data from DailyDialog on MELD. The performance gaps motivate us to investigate what could be causing the drop in performance.

		Test	
		MELD	DailyDialog
Train	MELD	**50.81**	35.89
	DailyDialog*	26.46	**40.60**
	DailyDialog	30.83	**55.04**

Table 2: Macro-F1 of emotion classification. *Average score of five randomly sampled sets of DailyDialog training data of equal size as MELD.

3.2. Linguistic Artifacts

	MELD	**DailyDialog**
Train Size	9,989	87,170
with TV-style	**1,391 (13.9%)**	**956 (1.1%)**
with Repetition	498 (5.0%)	90 (0.1%)
with Interjection	417 (4.2%)	486 (0.6%)
with Filler Words	589 (5.9%)	385 (0.4%)

Table 3: Statistics of linguistic style in MELD and DailyDialog training data

Manual examination of the datasets reveals a notable contrast between TV transcripts and everyday conversations in terms of the expression of emotions, where the former often include exaggerated or heightened expressions of emotion, while the later tend to be more subdued and realistic. We further observed a high frequency of repetition, interjection, and filler words, which appear much more frequently in TV transcripts (13.9%) than in daily conversations (1.1%), as shown in Table 3. As such, we considered these three elements part of the distinctive TV style:

Repetition We identified several types of repetition, including stuttering and repeating words or names to get the attention of other speakers.

Interjection MELD contains high-frequency of interjections such as "Oh!", "Wow!", "Oh God!" and "Ew!" convey strong emotions or serve to emphasize statements.

Filler Words Similar to Clark and Fox Tree (2002) and Dinkar et al. (2021), we also observe a prevalence of filler words, which are used to create a more relaxed or informal tone in TV shows.

The three identified TV-style features in TV transcripts could serve as relevant clues for the emotion recognition task, but such features are not as prevalent in everyday conversation. We hypothesize the ERC model, trained on TV transcripts, relies heavily on these specific linguistic artifacts to predict emotions, which may limit its ability to generalize effectively to daily conversational scenarios.

To validate our hypothesis, we re-trained the ERC model with a modified MELD train set stripped

Utterance	**Label**
Original: **Oh God,** this is so nerve wracking! **How-how** do you do this? **TV-style removed**: This is so nerve wracking! How do you do this?	Surprise
Original: **Umm, I-I** really don't want to tell this story. **TV-style removed**: I really don't want to tell this story.	Sadness

Table 4: Utterances in MELD with TV-style highlighted and removed.

		Test	
		MELD	DailyDialog
Train	Original	50.81	35.89
	TV-style removed	**46.72**	**38.02**

Table 5: Macro-F1 of emotion classification. Baseline models are trained with the original and TV-style removed MELD, respectively.

of these linguistic artifacts (Table 4). While the model's performance on MELD decreased by 4.1% in absolute terms when trained without these elements, we also observed a noteworthy 2.1% improvement on DailyDialog as seen in Table 5. This outcome provides evidence that in the absence of these linguistic artifacts, more general emotional features might be learned, leading to better adaptability.

4. Mitigation Strategies

We explore several changes to the model: 1) we use a contrastive loss to align the contextualized representations of utterances with and without the linguistic artifacts so that the model can better capture the similarity in emotional features while reducing the dependence of these artifacts, and 2) we introduce an auxiliary pseudo-emotion intensity regression loss to calibrate the emotion prediction.

4.1. Contrastive Learning

Given a conversation with N utterances, we feed the original conversation and a modified conversation with TV style removed through the hierarchical network of the baseline model described in Section 2 to obtain the contextualized representations for each utterance denoted as $[u_1, ..., u_N]$ and $[\hat{u}_1, ..., \hat{u}_N]$ respectively.

We employed the contrastive loss, InfoNCE (van den Oord et al., 2018) as follows:

$$L_c = -log\frac{exp(u \cdot \hat{u}_+/\tau)}{\sum_{i=0}^{K} exp(u \cdot \hat{u}_i/\tau)} \tag{1}$$

where τ is the temperature hyperparameter to pull the vector representations of the corresponding utterance pair closer.

4.2. Emotional Intensity

Interjections such as "Oh my god!" and "Wow!" in an utterance are strong indicators of the emotion "surprise". The intensity of the emotion expressed would be reduced without the interjection. Thus, we introduced a pseudo-emotion intensity score z_i for each utterance to reflect their emotional intensity. For an utterance that is labeled neutral, z_i is 0. For an emotional utterance with interjection removed, z_i is 0.5. For all original utterances with emotion, z_i is 1.0. We train a linear layer with a regression loss for z. Specifically, we used the mean squared error loss, which has the following formula:

$$L_{mse} = \frac{1}{U}\sum_{i}^{U}(z_i - \hat{z}_i)^2 \tag{2}$$

where $\hat{z}_i$ is the inferred intensity score of utterance u_i, and U is the total number of utterances per batch. To consider the inferred emotional intensity when making its final prediction, we use the intensity score to scale the emotion with the highest probability.

Together with the cross-entropy loss for emotion classification, we train the model end-to-end with contrastive and regression loss. The total loss for training is as follows:

$$L = L_{ce} + w_c L_c + w_r L_{mse} \tag{3}$$

where L_{ce} refer to the cross-entropy loss and the weights w_c and w_r are hyperparameters.

4.3. Experiment Settings

We perform a hyperparameter search for weights for contrastive and regression loss, τ, the temperature hyperparameter in contrastive loss (Equation 1) and the hidden size of the projection layer used in contrastive learning. We train the model with a batch size of 6, a linear learning rate schedule with 50 warm-up steps, a learning rate of $1e^{-5}$ for PLM weights, and $1e^{-3}$ for the rest of the model, on one

NVIDIA A6000 GPU with 48GB. The experiments were implemented using allennlp[1] and PyTorch[2].

5. Result and Analysis

Method	MELD	DailyDialog
Baseline	50.81	35.89
+ Contrastive Learning	48.04	40.18
+ Emotional Intensity	44.93	38.66
Proposed Method	**49.68**	**42.39**

Table 6: Macro-F1 of emotion classification. Models are trained with MELD training data and evaluated on MELD and DailyDialog test set.

As seen in Table 6, our proposed method achieved a significant absolute gain of 6.5% in performance on DailyDialog with a marginal drop of -1.1% in performance on MELD. It also outperformed the model trained with DailyDialog of equal size from Table 2 by 1.8%. The performance gain is evident in five of seven labels, as illustrated in Figure 1, demonstrating that our method has learned emotional features that generalize well to daily conversation.

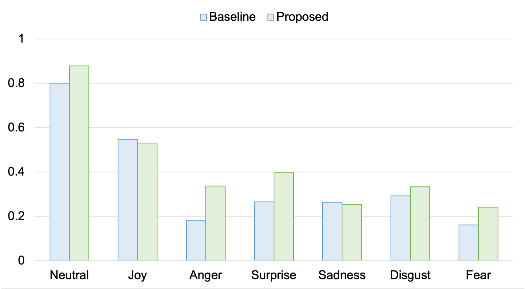

Figure 1: Performance on DailyDialog. F1 score for each label.

Ablation Studies Ablation results in Table 6 indicate that both techniques improve the model's generalization capabilities to daily conversations. However, only adding the regression task to the baseline model led to a significant drop in performance on MELD. We believe that the model overfits the regression task, leading to inaccurate scaling of the emotion prediction. Despite this, the two proposed losses are complementary and improve performance on non-TV conversations.

		Test	
		MELD	DailyDialog
Train	**TV-style removed**	**49.68**	**42.39**
	*Repetition removed**	48.44	36.61
	Interjection removed	46.83	39.18
	*Filler Words removed**	49.57	38.46

Table 7: Macro-F1 of emotion classification. Models are trained using the proposed method, and different TV-style elements are removed. *Only contrastive learning is used when repetition or filler words are removed.

In Table 7, we observe excluding interjections, which most likely contain emotional indicators, resulting in better performance than removing the other two characteristics. We hypothesize that when trained on the data with interjections removed, the model would have to rely more heavily on other subtle cues to infer the speaker's emotional state. This suggests that the proposed technique learns to capture other emotional features in the data beyond interjections.

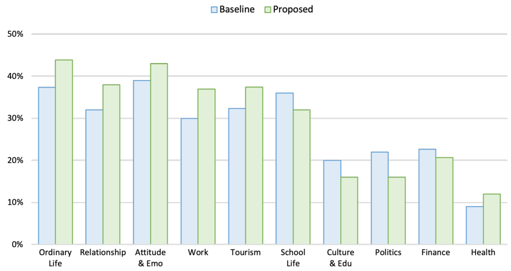

Figure 2: Performance on DailyDialog. F1 score for each topic.

We also dive deeper to analyze how our proposed method improves the classification performance on a range of daily conversation topics available in the DailyDialog dataset. The topics in Figure 2 are sorted according to the size. With sufficient examples, the proposed techniques are effective in learning generic emotional features on diverse daily conversation topics.

6. Limitations

Identifying all stylistic differences between TV transcripts and daily conversations based on lexical features is challenging. While we only identified three differences, other stylistic differences and information beyond the conversation, like culture or commonsense, could also explain the generalization gap. Furthermore, our model is trained on the TV series Friends, which may not represent diverse perspectives and experiences. This could lead to biases in the model's understanding and recognition of emotion. Therefore, knowing these potential limitations and biases is crucial when using the proposed model for real-world applications.

[1] https://allenai.org/allennlp
[2] https://pytorch.org/

7. Conclusion

In this paper, we highlight the limitations of current ERC models trained on the TV series dataset when applied to the daily conversation dataset. Our proposed approach demonstrates a first step in effectively addressing these limitations by mitigating linguistic artifacts and emphasizing emotional features. Our study on adaptability shows potential for developing a robust ERC model that can be utilized in real-world applications.

Acknowledgments

This research is partially supported by the programme DesCartes funded by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme.

8. Bibliographical References

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Herbert H. Clark and Jean E. Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Tanvi Dinkar, Beatrice Biancardi, and Chloé Clavel. 2021. From local hesitations to global impressions of a speaker's feeling of knowing. In *Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, pages 232–241, Trento, Italy. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481, Online. Association for Computational Linguistics.

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.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Joosung Lee and Wooin Lee. 2022. CoMPM: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5669–5679, Seattle, United States. Association for Computational Linguistics.

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. *ArXiv*, abs/1907.11692.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883, Vancouver, Canada. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. 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.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748.

Sayyed Zahiri and Jinho D. Choi. 2018. Emotion Detection on TV Show Transcripts with Sequence-based Convolutional Neural Networks. In *Proceedings of the AAAI Workshop on Affective Content Analysis*, AFFCON'18, pages 44–51, New Orleans, LA.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. 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 165–176, Hong Kong, China. Association for Computational Linguistics.