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 (I2R), A*STAR, Singapore

*Work done when Donovan was working at I2R

Emotion Recognition for Conversations

- To identify the emotion expressed by the speaker for each utterance

- Has many potential applications, including improving customer service, enhancing personal relationships, and diagnosing and treating mental health conditions.

Existing Works

- Focus on training and testing models on the same datasets, and **there is no prior work on adaptability**

- Hindered by the challenges of unifying datasets with **different emotion taxonomies and conversation settings**, including TV series, daily conversations, and social media

Adaptability of ERC models

- aims to address this knowledge gap by presenting a preliminary investigation into the adaptability of ERC models

- 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' reli on these artifacts.

Adaptability Study - Methodology

- MELD (Poria et al., 2019) and DailyDialog (Li et al., 2017)
    - Both employ the same set of emotion labels (joy, anger, sadness, fear, disgust, surprise, and neutral).

- Evaluation Metric: Macro-F1
- Baseline: RoBERTa + LSTM

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%

Adaptability Study - 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.

Significant performance gap for out-of-distribution dialogs.

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

Mitigation Strategies

1. Contrastive Learning
   - Pull the vector representations of the pair of utterances with and without TV styles closer

1. Emotional Intensity
   - Introduced a pseudo-emotion intensity score for each utterance to reflect their emotional intensity.
   - Train a linear layer to infer the intensity and scale the probability the emotion with th
     probability.

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.

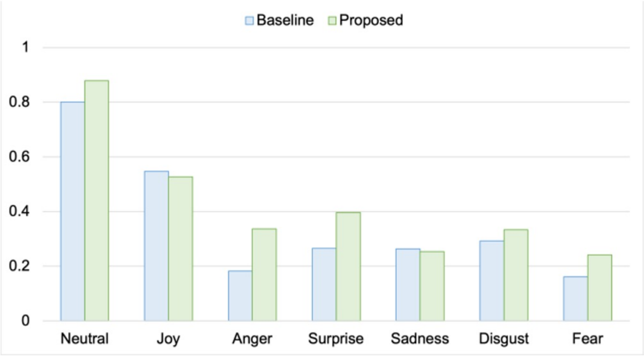

Figure 1: Performance on DailyDialog. F1 score for each label.

- Ablation: Training with only regression lead to overfitting

Ablation Studies

		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.

- Excluding interjections, which most likely contain emotional indicators, resulting in bet performance than removing the other two characteristics.

End