Sentiment Reasoning for Healthcare

Khai-Nguyen Nguyen*, Khai Le-Duc*, Bach Phan Tat, Duy Le, Long Vo-Dang, Truong Son Hy
knguyen07@wm.edu, duckhai.le@mail.utoronto.ca, TruongSon.Hy@indstate.edu

Introduction

- **Transparency in healthcare AI** is critical for decision-making and trust.
- Traditional healthcare sentiment analysis lacks reasoning and explainability
- We propose **Sentiment Reasoning**, a novel task that integrates rationale generation into sentiment classification

Contributions

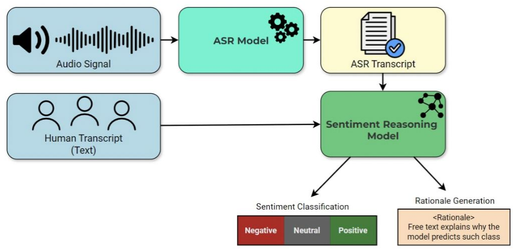
- New task: Sentiment Reasoning for speech and text modalities.
- Developed MultiMed-SA, a sentiment reasoning dataset for medical conversations, and a multimodal speech-text Sentiment Reasoning framework
- Provide in-depth analysis of rationale / Chain-of-Thought (CoT)-augmented training

MultiMed-SA

Split	Label	Count	Percentage
Train	Neutral	2844	49.94%
	Negative	1694	29.74%
	Positive	1157	20.32%
Test	Neutral	958	43.88%
	Negative	701	32.11%
	Positive	524	20.01%

ENG Translation	Label	Rationale
The patient will suffer from emotional disorder and sometimes depression	NEG.	Emotional disorder
Stroke is related to the formation of blood clots and the fact that these blood clots travel	NEG.	Negative medical condition
It's often confused with antiplatelet drugs	NEG.	Confusion
A crucial point is that the overweight patient	NEU.	Sharing advice
The cortisol hormone in blood as well as catecholamine	NEU.	Objective description of hormones
You could call these blood-thinning drugs or other names, and it can	NEU.	Objective description
It is not expensive, luckily, in recent years there are another group of medicine	POS.	Expressing luck
To reduce and eliminate the formation of these blood clots, we use several measures, one of which is	POS.	Avoid forming blood clots
This group of drugs has been around for a very long time and is very cheap, with no cost	POS.	Long-standing and inexpensive medication

Sentiment Reasoning



Consists of 2 Sub-tasks:

1. **Sentiment Classification**: Predict sentiment labels
2. **Rationale Generation**: Generate explanations for predictions.

Training with Rationale

- **Multitask Training**: CoT-augmented tasks for encoder-decoders.
- **Post-Thinking**: Rationale appended to training targets for decoders.

Results on Human Transcripts

Model	Acc.	F1 Neg.	F1 Neu.	F1 Pos.	Mac F1	R-1	R-2	R-L	R-Lsum	BERTscore
Encoder (Label Only)										
PhoBERT	0.6674	0.6969	0.6607	0.6377	0.6651					
ViHealthBERT	0.6752	0.6970	0.6718	0.6535	0.6741					
Encoder-Decoder (Label Only)										
ViT5	0.6628	0.6922	0.6687	0.6007	0.6545					
BARTpho	0.6523	0.6870	0.6571	0.5841	0.6427					
Decoder (Label Only)										
vmlu-llm	0.6592	0.6768	0.6769	0.5911	0.6483					
Vistral7B	0.6716	0.6858	0.6771	0.6398	0.6676					
Encoder-Decoder (Label + Rationale)										
ViT5	0.6633	0.6936	0.6572	0.6335	0.6615	0.3910	0.2668	0.3653	0.3660	0.8093
BARTpho	0.6619	0.7029	0.6460	0.6265	0.6585	0.3871	0.2613	0.3658	0.3683	0.8077
Decoder (Label + Rationale)										
vmlu-llm	0.6729	0.7039	0.6714	0.6307	0.6687	0.3947	0.2467	0.3789	0.3796	0.8086
Vistral7B	0.6812	0.7152	0.6765	0.6425	0.6781	0.4155	0.2788	0.3880	0.3900	0.8101

Table 2: Baseline performance of encoders, encoder-decoders, and decoders on the Vietnamese human transcript. From left to right is: Accuracy, F1-{negative, neutral, positive, macro}, ROUGE-{1, 2, L, Lsum}, BERTscore. The **Label Only** models are models trained only with the label, serving as the baseline, while **Label + Rationale** indicates models trained with rationale. As the **Label Only** models are not trained to generate rationale, we do not evaluate them on ROUGE and BERTscore.

Results on ASR Transcripts

Model	Acc.	F1 Neg.	F1 Neu.	F1 Pos.	Mac F1	R-1	R-2	R-L	R-LSum	BERTscore
Encoder (Label Only)										
PhoBERT	0.6166	0.6418	0.6231	0.5658	0.6102					
ViHealthBERT	0.6198	0.6307	0.6261	0.5934	0.6167					
Encoder-Decoder (Label Only)										
ViT5	0.6157	0.6412	0.6258	0.5523	0.6064					
BARTpho	0.6056	0.6364	0.6156	0.5311	0.5944					
Decoder (Label Only)										
vmlu-llm	0.6216	0.6296	0.6551	0.5186	0.6011					
Vistral7B	0.6299	0.6377	0.6537	0.5609	0.6174					
Encoder-Decoder (Label + Rationale)										
ViT5	0.6189	0.6305	0.6286	0.5837	0.6143	0.3571	0.2202	0.3350	0.3366	0.8044
BARTpho	0.6129	0.6523	0.6028	0.5665	0.6072	0.3956	0.2652	0.3728	0.3774	0.8106
Decoder (Label + Rationale)										
vmlu-llm	0.6395	0.6585	0.6557	0.5723	0.6289	0.3853	0.2386	0.3663	0.3671	0.8092
Vistral7B	0.6354	0.6485	0.6479	0.5892	0.6285	0.3558	0.2237	0.3343	0.3394	0.7994

Table 3: Baseline performance of encoders, encoder-decoders, and decoders on the Vietnamese ASR transcript. Further information about our metrics can be found in Table 2.

Rationale Evaluation

Model	Acc.	F1 Neg.	F1 Neu.	F1 Pos.	Mac F1
Encoder-Decoder (Label + Rationale)					
ViT5_human	0.6633	0.6936	0.6572	0.6335	0.6615
ViT5_elaborate	0.6661	0.6903	0.6799	0.5985	0.6562
ViT5_cot	0.6619	0.6968	0.6552	0.6237	0.6586
BARTpho_human	0.6619	0.7029	0.6460	0.6265	0.6585
BARTpho_elaborate	0.6564	0.7031	0.6528	0.5870	0.6476
BARTpho_cot	0.6464	0.6922	0.6611	0.5287	0.6273
Decoder (Label + Rationale)					
Vistral7B_human	0.6812	0.7152	0.6765	0.6425	0.6781
Vistral7B_elaborate	0.6688	0.6846	0.6647	0.6564	0.6685
Vistral7B_cot	0.6706	0.6725	0.6807	0.6477	0.6670
vmlu-llm_human	0.6729	0.7039	0.6714	0.6307	0.6687
vmlu-llm_elaborate	0.6867	0.7203	0.6868	0.6353	0.6808
vmlu-llm_cot	0.6821	0.6966	0.6779	0.6711	0.6819

Table 4: Performance of generative models on the different rationale formats on our test set. Human/elaborate/CoT specifies the format of rationale the model was trained on.

Key takeaways

1. Encoders are efficient yet effective sentiment classification baselines

2. ASR errors (WER 29.6%) have a marginally negative impact on sentiment classification

3. Rationale-augmented training improve model performance

4. The format of post-thinking rationale doesn't affect the generative models performance

5. Models are likely to misclassify *POSITIVE* and *NEGATIVE* transcripts as *NEUTRAL*

6. Generated rationales have different vocabulary to that of human but with similar semantics

7. No significant difference in the semantic quality of generated rationales between human and ASR transcripts

References