Counterfactual Reasoning for Out-of-distribution Multimodal Sentiment Analysis

Teng Sun†, Wenjie Wang§, Liqiang Jing†,
Yiran Cui†, Xuemeng Song†, Liqiang Nie†

†Shandong University, Shandong, China,
§National University of Singapore, Singapore, Singapore

Presenter: Liqiang Jing

jingliqiang6@gmail.com

Outline

Background

Related Work

Motivations

Method

Experiment

Background

Users tend to present their opinions on social media platforms.

How Many Tweet Sent Per Day of 2013-2022

Years	Average Number of Tweets (in million)
2012	340
2013	500
2014	546
2015	592
2016	634
2017	683
2018	729
2019	775
2020	821
2021-2022	867

https://www.renolon.com/number-of-tweets-per-day/

Background

- Sentiment Analysis

Derek Jeter
@derekjeter

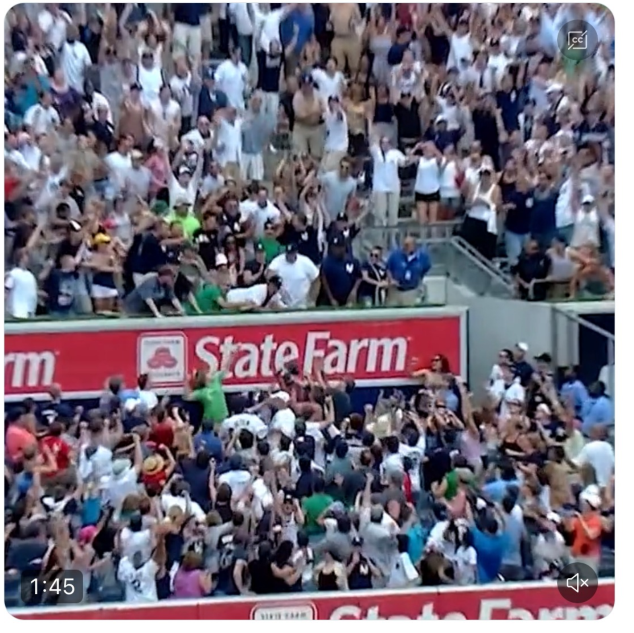
Being named captain of the New York Yankees was one of the greatest honors of my career. #TheCaptain

Textual Tweet

.@derekjeter's 3,000th hit game was iconic

#TheCaptain

翻译推文



Multimodal Tweet

Task Definition

- Multimodal Sentiment Analysis (MSA)

  Predict the sentiment label based on three modalities (i.e., text, video, and audio).

Outline

- Background
- Related Work
- Motivations
- Method
- Experiment

Related Work

MSA

- Multimodal representation learning
  - Translation-based
  - Correlation-based
  - Shared subspace learning
- Cross-modal fusion
  - Feature-level fusion
  - Decision-level fusion

Outline

Background

Related Work

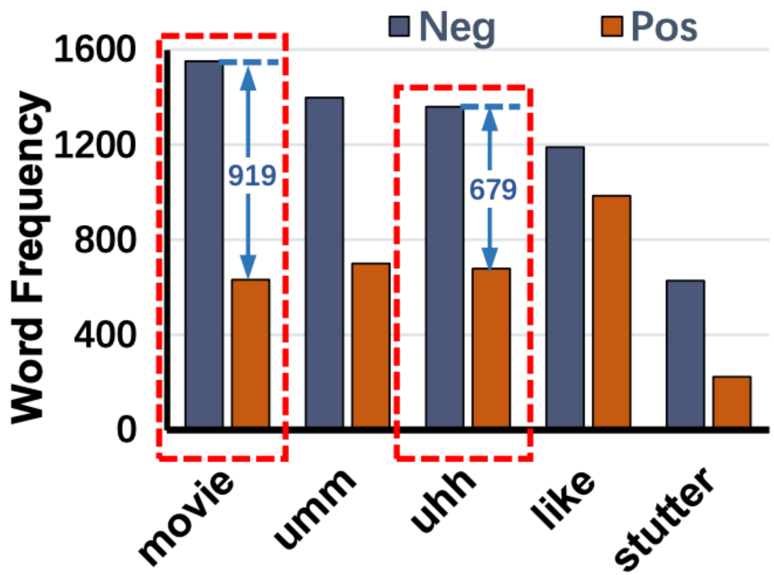
Motivations

Method

Experiment

Motivations

➢ Existing studies usually suffer from fitting the spurious correlations in textual modality.

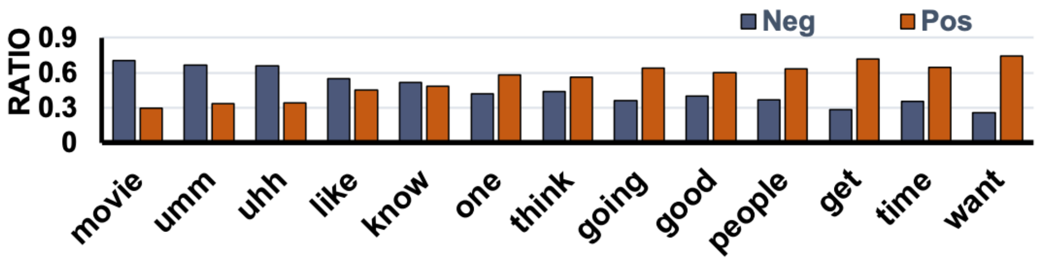


Distribution of the most frequent words in MOSEI dataset.

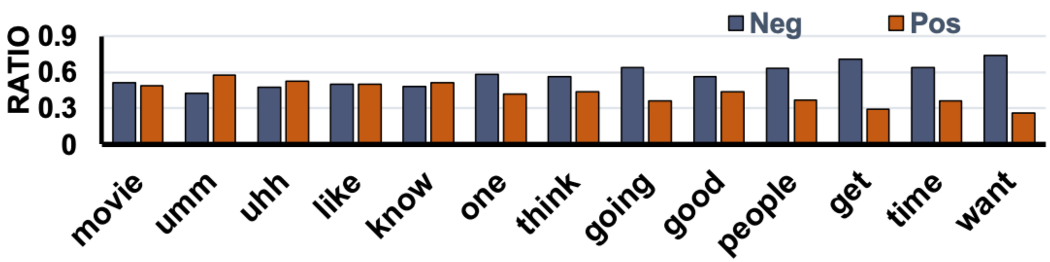
Task Definition

➢ OOD MSA Task

Construct the OOD testing set for each biased dataset, with significantly different word-sentiment correlations from the training one.



(a) Distribution of the most frequent words in the training set.



(b) Distribution of the most frequent words in the OOD testing set.

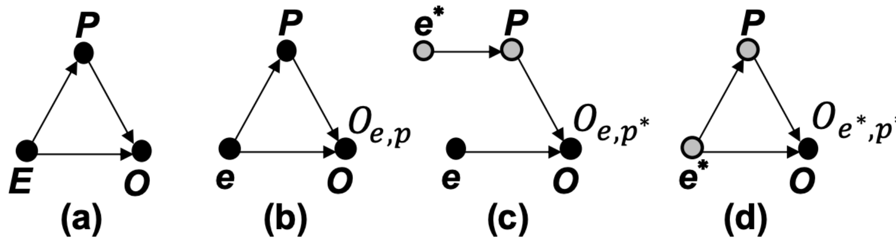
Keys

- Disentangling the good and bad effects of textual modality on the model prediction.
- Mitigating the bad effect for stronger out-of-distribution (OOD) generalization.
- Utilizing multimodal cues to alleviate the textual correlations.

Outline

- Background
- Related Work
- Motivations
- Method
- Experiment

Preliminary



The causal graph of the admission outcome, where the admission outcome of graduate (O) is directly affected by the experience (E) and publication (P).

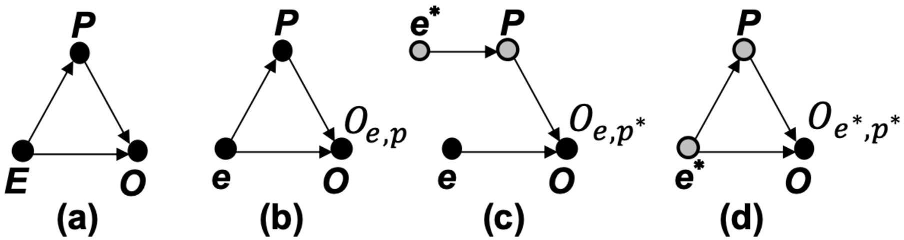
- Structural Equations

$$P_e = p = f_P(E = e), O_{e,p} = f_O(E = e, P = p),$$

- Total Effect

$$\begin{cases} \mathrm{TE} = O_{e,p} - O_{e^*,p^*} = f_O(E = e, P = p) - f_O(E = e^*, P = p^*), \\ p^* = P_{e^*} = f_P(E = e^*), \end{cases}$$

Preliminary



The causal graph of the admission outcome, where the admission outcome of graduate (O) is directly affected by the experience (E) and publication (P).

- Natural Direct Effect

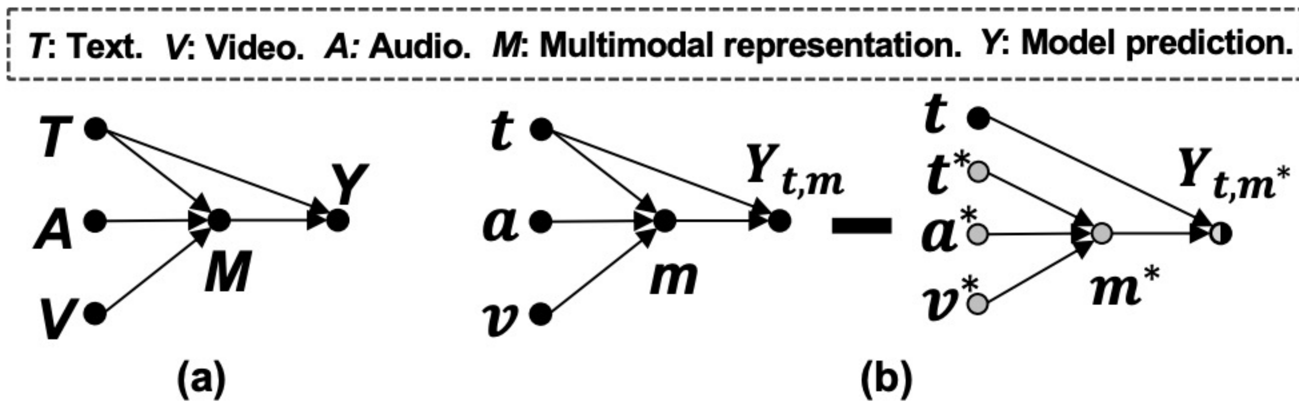
$$\mathrm{NDE} = O_{e,p^*} - O_{e^*,p^*},$$

- Total Indirect Effect

$$\mathrm{TE} = \mathrm{NDE} + \mathrm{TIE}.$$

$$\mathrm{TIE} = \mathrm{TE} - \mathrm{NDE} = O_{e,p} - O_{e,p^*}.$$

Causal Graph of MSA



(a) The causal graph in the MSA. (b) The illustration of counterfactual inference.

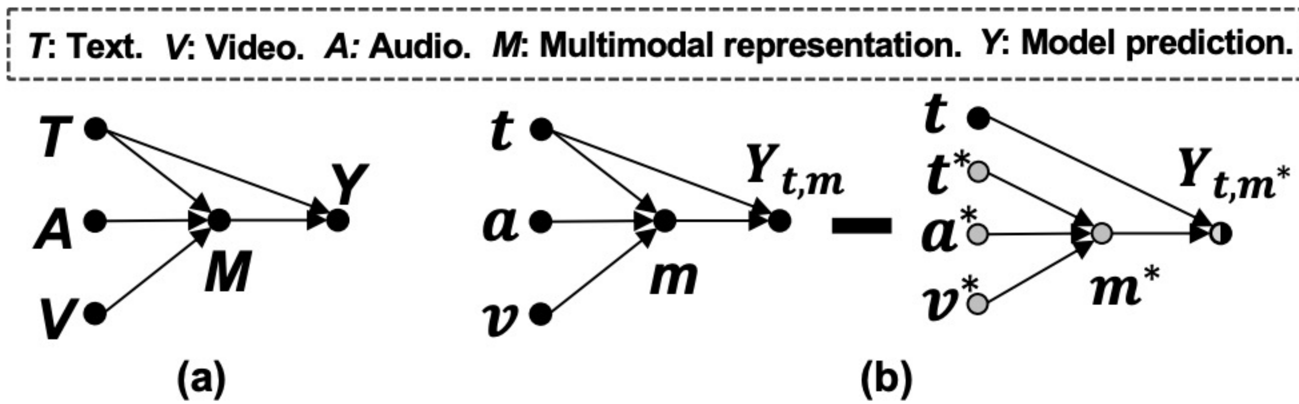
- Casual Relationships

$$\begin{cases} Y_{t,m} = f_Y(T = t, M = m), \\ m = f_M(T = t, A = a, V = v). \end{cases}$$

- TE of Textual Modality

$$\mathrm{TE} = Y_{t,m} - Y_{t^*,m^*} = f_Y(T = t, M = m) - f_Y(T = t^*, M = m^*),$$

Causal Graph of MSA



(a) The causal graph in the MSA. (b) The illustration of counterfactual inference.

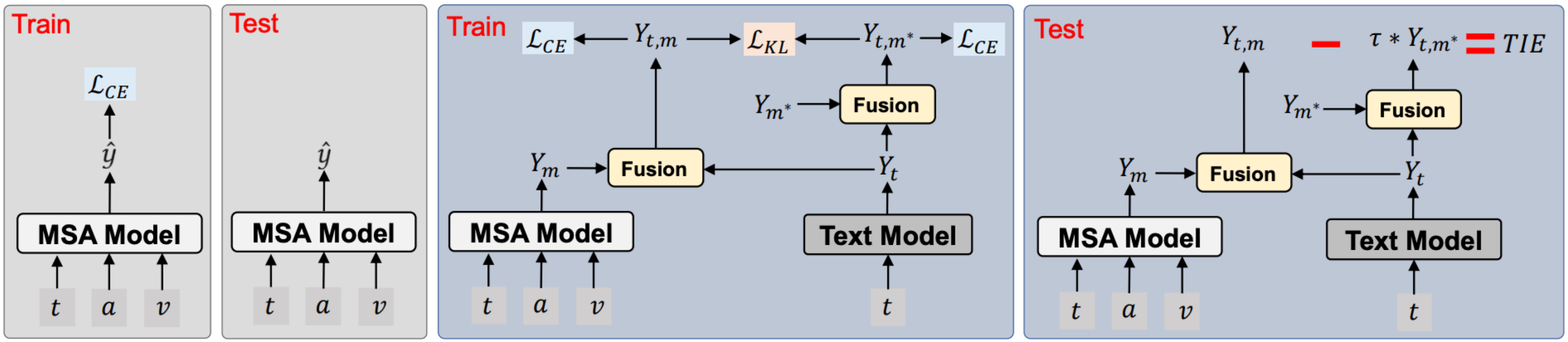
- NDE of Textual Modality

$$Y_{t,m^*} - Y_{t^*,m^*} = f_Y(T = t, M = m^*) - f_Y(T = t^*, M = m^*),$$

- TIE of Textual Modality

$$\mathrm{TIE} = \mathrm{TE} - \mathrm{NDE} = Y_{t,m} - Y_{t,m^*},$$

CLUE



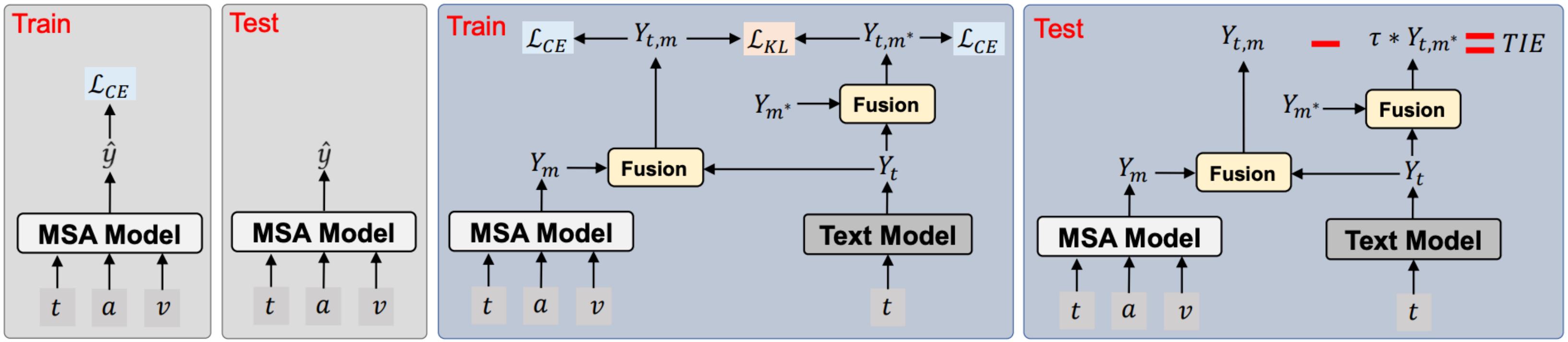
(a) Traditional MSA. (b) Counterfactual inference of CLUE.

- Implementation

$$\boxed{Y_m = f_M(T = t, A = a, V = v)}$$

$$\boxed{Y_{t,m} = f_Y(T = t, M = m) = h(Y_t, Y_m) = \mathrm{SUM}(Y_t, Y_m) = \log \sigma(Y_t + Y_m),}$$

CLUE



(a) Traditional MSA.

(b) Counterfactual inference of CLUE.

- Training

$$\mathcal{L}_{CE} = \alpha * \mathrm{CE}(Y_{t,m}, y) + \beta * \mathrm{CE}(Y_{t,m^*}, y),$$

- Testing

$$\mathrm{TIE} = Y_{t,m} - \tau * Y_{t,m^*} = h(Y_t, Y_m) - \tau * h(Y_t, Y_{m^*}),$$

Outline

Background

Related Work

Motivations

Method

Experiment

ODD Dataset Construction

Algorithm 1 IID and OOD Set Construction.

Input: The whole dataset $\mathcal{D}$, the pre-defined distribution difference $\boldsymbol{\phi}_\Delta$, the number of iterations n, simulated annealing temperature τ, and the temperature decay rate α.

Output: IID set $\mathcal{D}_{iid}$ and OOD set $\mathcal{D}_{ood}$.

```
Get an IID set $\mathcal{D}_{iid}$ and OOD set $\mathcal{D}_{ood}$ by random splitting $\mathcal{D}$.
Compute distributions $\boldsymbol{\phi}_{iid}$ and $\boldsymbol{\phi}_{ood}$ of all words over different sentiment categories in $\mathcal{D}_{iid}$ and $\mathcal{D}_{ood}$, respectively.
Set $V = \|abs(\boldsymbol{\phi}_{iid} - \boldsymbol{\phi}_{ood}) - abs(\boldsymbol{\phi}_\Delta)\|_1$.
repeat
    repeat
        Randomly swap samples between $\mathcal{D}_{iid}$ and $\mathcal{D}_{ood}$ by perturbation strategies[2] to a new IID set $\hat{\mathcal{D}}_{iid}$ and OOD set $\hat{\mathcal{D}}_{ood}$.
        Calculate $\hat{\boldsymbol{\phi}}_{iid}$ ($\hat{\boldsymbol{\phi}}_{ood}$) with $\hat{\mathcal{D}}_{iid}$ ($\hat{\mathcal{D}}_{ood}$), respectively.
        Set $\hat{V} = \|abs(\hat{\boldsymbol{\phi}}_{iid} - \hat{\boldsymbol{\phi}}_{ood}) - abs(\boldsymbol{\phi}_\Delta)\|_1$.
        Get a random number $R$ and $0 \le R < 1$.
        if $V \ge \hat{V}$ then
            Set $\mathcal{D}_{iid}, \mathcal{D}_{ood}, V = \hat{\mathcal{D}}_{iid}, \hat{\mathcal{D}}_{ood}, \hat{V}$.
        else if $exp((\hat{V} - V)/\tau) > R$ then
            Set $\mathcal{D}_{iid}, \mathcal{D}_{ood}, V = \hat{\mathcal{D}}_{iid}, \hat{\mathcal{D}}_{ood}, \hat{V}$.
        end if
    until Swapping times reach $n$.
    Set $\tau = \tau * \alpha$.
until Iteration times reach $n$.
```

Experiment

➢ On Model Comparison

Table 1: OOD testing performance (%) comparison among different methods on MOSEI and MOSI datasets. For *Acc-2* and *F1*, "∗" is calculated as "negative/non-negative" and "§" is calculated as "negative/positive". The best result of each pair of the original MSA model and the model with CLUE is highlighted in bold.

Model	MOSEI					MOSI				
	2-class				7-class	2-class				7-class
	*Acc-2**	*F1**	*Acc-2*§	*F1*§	*Acc-7*	*Acc-2**	*F1**	*Acc-2*§	*F1*§	*Acc-7*
TFN	71.23	70.46	69.76	69.02	41.05	73.02	72.93	74.62	74.56	32.95
LMF	68.16	68.31	69.58	69.58	31.11	73.54	73.40	75.27	75.18	29.10
MulT	72.56	72.44	73.73	73.58	40.58	75.00	74.75	76.72	76.52	29.80
MAG-BERT	74.59	74.48	76.41	76.27	45.88	75.57	75.52	77.28	77.26	39.85
+CLUE (Ours)	**78.34**$^{+3.75}$	**78.23**$^{+3.75}$	**80.51**$^{+4.10}$	**80.46**$^{+4.19}$	**48.66**$^{+2.78}$	**77.25**$^{+1.68}$	**77.46**$^{+1.94}$	**78.65**$^{+1.37}$	**78.83**$^{+1.57}$	**40.75**$^{+0.90}$
MISA	74.48	74.39	76.45	76.33	43.15	75.90	75.82	77.39	77.35	38.05
+CLUE (Ours)	**77.17**$^{+2.69}$	**77.08**$^{+2.69}$	**78.77**$^{+2.32}$	**78.74**$^{+2.41}$	**46.86**$^{+3.71}$	**78.25**$^{+2.35}$	**78.28**$^{+2.46}$	**79.17**$^{+1.78}$	**79.19**$^{+1.84}$	**42.25**$^{+4.20}$
Self-MM	74.68	74.33	74.50	74.22	45.81	76.70	76.68	78.12	78.13	40.25
+CLUE (Ours)	**77.76**$^{+3.08}$	**77.72**$^{+3.39}$	**79.48**$^{+4.98}$	**79.47**$^{+5.25}$	**48.09**$^{+2.28}$	**78.75**$^{+2.05}$	**78.75**$^{+2.07}$	**79.94**$^{+1.82}$	**79.93**$^{+1.80}$	**41.75**$^{+1.50}$

CLUE consistently surpasses all the baselines, exhibiting the effectiveness of the proposed scheme.

Experiment

➢ On Model Comparison

Table 2: IID testing performance (%) comparison among different methods on MOSEI and MOSI datasets. *Acc-2* and *F1* are calculated as "negative/non-negative". We omitted the similar results of "negative/positive" to save space.

Model	MOSEI			MOSI		
	2-class		7-class	2-class		7-class
	Acc-2	*F1*	*Acc-7*	*Acc-2*	*F1*	*Acc-7*
TFN	81.59	81.54	52.11	80.24	80.31	40.07
LMF	79.59	80.34	48.27	79.85	79.95	35.04
MulT	81.05	81.44	53.21	79.61	79.71	35.19
MAG-BERT	82.82	83.19	53.52	83.91	83.96	46.97
+CLUE (Ours)	**84.62**	**85.46**	**53.68**	**84.37**	**84.28**	**48.84**
MISA	82.17	82.61	**53.26**	83.52	83.58	45.26
+CLUE (Ours)	**84.51**	**85.28**	53.15	**84.07**	**84.16**	**46.31**
Self-MM	83.71	83.80	53.31	84.14	84.17	**48.74**
+CLUE (Ours)	**84.52**	**84.46**	**53.42**	**84.31**	**84.38**	48.04

CLUE consistently surpasses all the baselines, exhibiting the effectiveness of the proposed scheme.

Experiment

➢ On Ablation Study

Table 3: Ablation study results (%) for the binary classification (negative/non-negative) of our proposed CLUE on MOSEI. The best results are highlighted in boldface.

Model	IID testing		OOD testing	
	Acc-2	*F1*	*Acc-2*	*F1*
MAG-BERT+CLUE	84.62	85.46	**78.34**	**78.23**
w/o-MSA model	80.31	81.52	65.49	68.01
w/o-text model	**85.09**	**85.60**	74.37	74.83
w/o-KL loss	84.55	84.45	78.28	78.17
MISA+CLUE	84.51	85.28	**77.17**	**77.08**
w/o-MSA model	80.31	81.52	65.49	68.01
w/o-text model	84.31	85.26	74.53	75.78
w/o-KL loss	**84.69**	**85.44**	76.75	76.77
Self-MM+CLUE	84.52	84.46	**77.76**	**77.72**
w/o-MSA model	80.31	81.52	65.49	68.01
w/o-text model	84.52	**85.41**	73.51	74.12
w/o-KL loss	**84.63**	84.53	77.41	77.43

CLUE obtains the best performance, which verifies these components are significant in our model.

Conclusion

- We define a novel **OOD** MSA task, which points out the **spurious correlations** in textual modality and highlights the necessity of strong OOD **generalization** abilities.

- We devise a **model-agnostic CLUE** framework. It strengthens the existing MSA models via capturing **the causal relationships** in the training set and **mitigating the bad effect** of textual modality by the counterfactual inference.

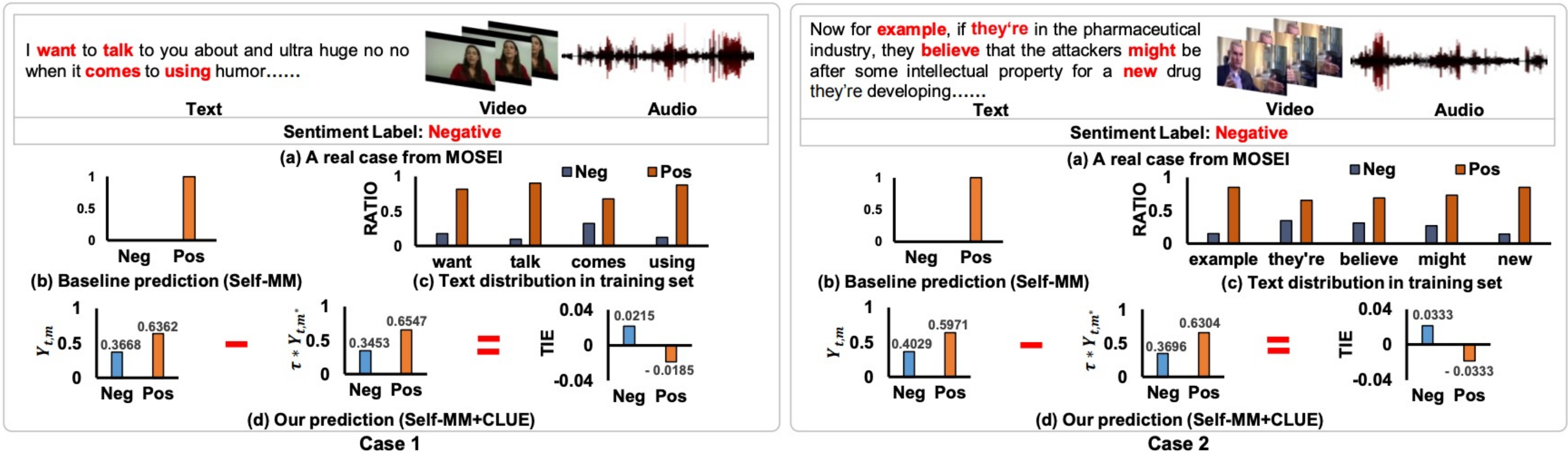
- We conduct extensive experiments on two benchmark datasets, and the results demonstrate the **superior effectiveness and generalization ability** of CLUE.

Thanks for your listening.

Codes are available!

Experiment

➢ On Case Study



Cases of the binary classification by self-MM and CLUE.