SALAD: Improving Robustness and Generalization through Contrastive Learning with Structure-Aware and LLM-Driven Augmented Data

Suyoung Bae[1] , Hyojun Kim[2] , YunSeok Choi[1*], Jee-Hyong Lee[1*]
[1]Sungkyunkwan University, [2]SK Telecom
[1]{sybae01, ys.choi, john}@skku.edu [2]hjkim@sk.com

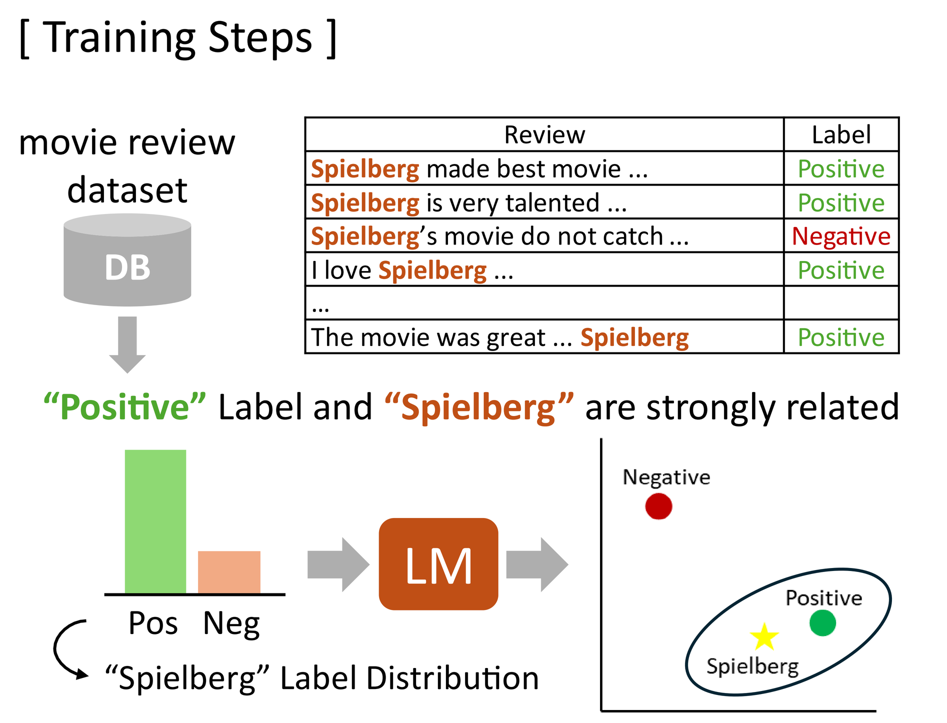
NAACL 2025

1. Problem: Spurious Correlations in NLP tasks

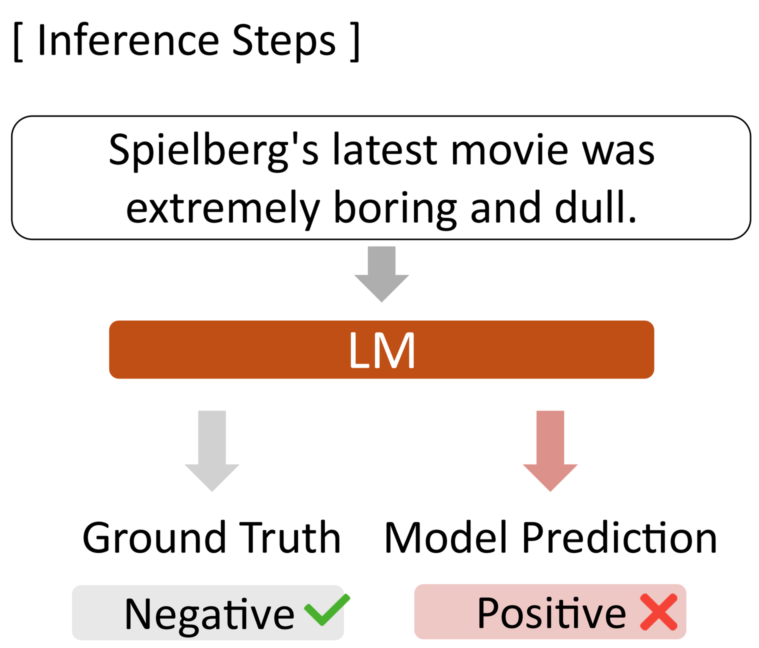
- Spurious correlation occurs when some variable and label appear strongly related, but there's no genuine causal relationship.

➢ **Scenario:** When we use a movie review dataset to perform a sentiment analysis task, where the dataset frequently mentions the famous director **"Spielberg"** in positive contexts.

Reason: Training dataset unbalancing



Result: Spurious Correlation



2. Task & Overview

Task Objective:

Effectively reduce spurious correlation in various NLP tasks using contrastive learning without any additional dataset

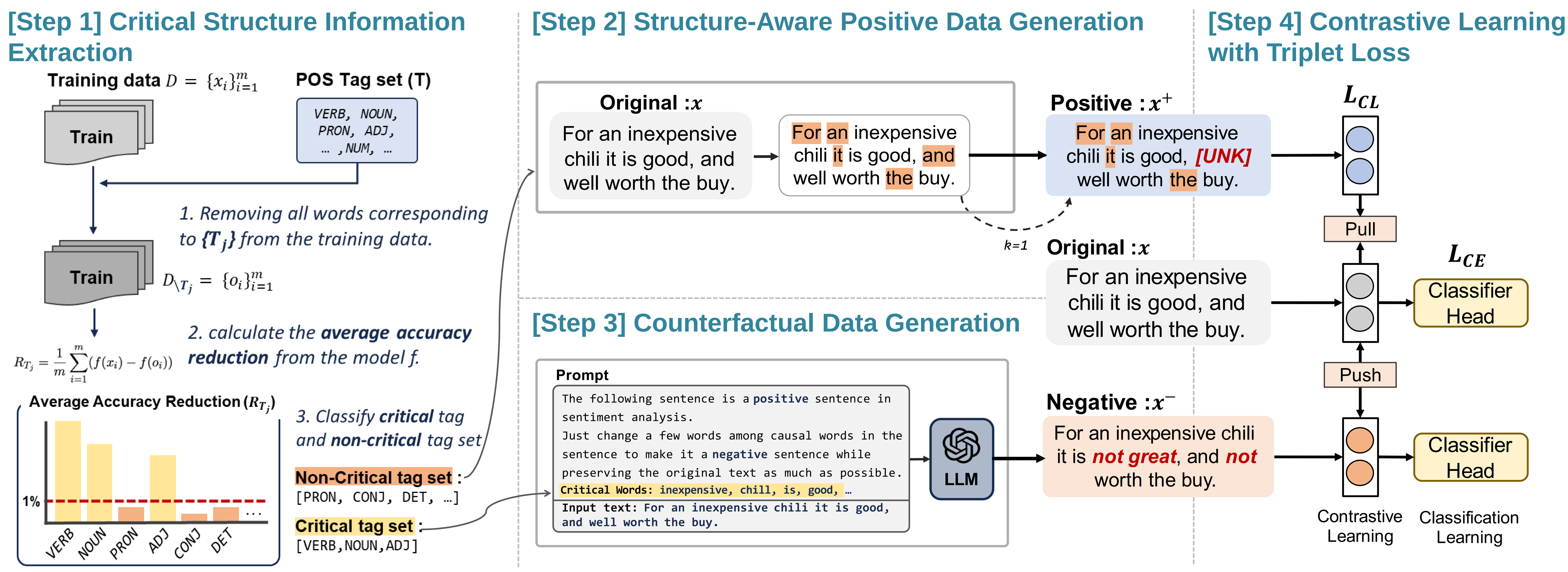
Overview:

[1] Extracting **critical & non-critical** structures in each task
[2] Using **non-critical** structures to generate positive data
[3] Using **critical** structures to generate negative data
[4] Contrastive Learning for effective training

Q. What is the **critical structures** where shortcut occurs?
A. Some **critical POS tags** influence the label, but some are not

	Sentence	Label
Original sentence	Spielberg's latest movie was extremely boring and dull. NOUN ADJ NOUN VERB ADB ADJ CC ADJ	**Negative**
Changing critical tags	Spielberg's latest movie was extremely exciting and fun. ADJ ADJ	**Positive** (Changed)
Changing non-critical tags	Spielberg's latest movie is extremely boring and dull. VERB	**Negative** (Not Changed)

3. Proposed Method: SALAD



4. Experiment Results

[Table 1] Task1: Sentiment Classification Task

Methods	In-Domain Dataset		Out-of-Distribution Dataset				Overall
	O-Test	CF-Test	YELP	SST2	FindFood	Tweet	
Standard Fine-Tuning (full-data)							
RoBERTa-large (Liu et al., 2019)	**94.13**	92.28	94.85	79.41	95.24	73.04	88.16
Robust Learning							
SupCon (Gunel et al., 2021)	93.85	88.11	95.26	86.20	95.32	74.90	88.94
C2L (Choi et al., 2022)	93.37	93.03	93.19	79.90	94.26	68.85	87.10
Text Data Augmentation							
EDA (Wei and Zou, 2019)	93.58	93.72	95.28	89.73	95.40	81.24	91.49
SSMBA (Ng et al., 2020)	93.60	92.69	**95.90**	89.40	**96.12**	78.75	91.08
AugGPT (Dai et al., 2023)	93.37	91.46	95.32	90.21	94.18	78.66	90.53
Counterfactual Data Augmentation							
Human-CAD (Kaushik et al., 2020)	93.17	95.47	92.16	88.65	94.26	80.66	90.73
CORE-CAD (Dixit et al., 2022)	91.73	95.15	89.70	90.10	93.06	**86.77**	91.09
SALAD	93.78	**95.90**	94.99	**92.68**	95.58	85.35	**93.05**

[Table 3] Task 3: Natural Language Inference

Methods	In-Domain		Out-of-Distribution		Overall
	O-test	CF-test	MNLI[1]	MNLI[2]	
Standard Fine-Tuning (full-data)					
RoBERTa-large (Liu et al., 2019)	87.50	69.90	73.27	73.97	76.16
Robust Learning					
SupCon (Gunel et al., 2021)	86.42	60.03	64.70	64.39	68.89
C2L (Choi et al., 2022)	87.96	68.49	72.18	72.74	75.34
Text Data Augmentation					
EDA (Wei and Zou, 2019)	86.59	67.58	70.93	71.12	74.06
SSMBA (Ng et al., 2020)	87.16	63.54	72.03	72.95	73.92
AugGPT (Dai et al., 2023)	86.92	69.61	73.62	74.38	76.13
Counterfactual Data Augmentation					
Human-CAD (Kaushik et al., 2020)	88.25	71.60	71.74	71.47	75.76
CORE-CAD (Dixit et al., 2022)	64.65	57.26	62.60	62.98	61.88
DISCO (Chen et al., 2023)	79.84	78.66	68.42	67.60	73.63
SALAD	**88.40**	**80.91**	**74.06**	**74.93**	**79.57**

[Table 2] Task2: Sexism Classification

Methods	IDD		ODD	Overall
	O-Test	CF-Test	Tweet	
RoBERTa-large	92.69	49.23	81.00	72.49
SupCon	91.79	22.56	76.28	60.84
C2L	**93.21**	37.69	77.92	67.18
EDA	91.67	37.69	81.59	67.74
SSMBA	92.82	25.64	79.36	63.02
AugGPT	92.31	29.23	78.83	64.08
Human-CAD	91.79	**91.80**	83.11	**89.47**
SALAD	93.07	88.47	**83.38**	88.31

[Table 4] Cross-domain Task

Methods	S → I	S → F	I → S	I → F	F → S	F → I	Overall
Standard Fine-Tuning (full-data)							
RoBERTa-large (Liu et al., 2019)	91.67	93.08	89.16	91.13	82.48	90.22	89.62
Robust Learning							
SupCon (Gunel et al., 2021)	90.82	89.64	91.21	94.95	73.40	89.68	88.28
C2L (Choi et al., 2022)	90.52	91.61	89.90	94.64	81.18	90.50	89.72
Text Data Augmentation							
EDA (Wei and Zou, 2019)	91.64	93.51	90.76	94.12	80.18	89.29	89.92
SSMBA (Ng et al., 2020)	90.71	90.78	**94.21**	93.96	78.75	89.31	89.62
SALAD	**92.41**	**94.19**	90.88	**94.96**	**86.00**	**91.25**	**91.61**

5. Conclusions

1. **Improved training robustness** by enabling the model to learn structural patterns and apply contrastive learning.
2. **Achieved generalizability** by performing well on out-of-distribution domains.
3. **Ensured consistent performance** across a variety of sentence structures by enabling the model to learn structural patterns where shortcuts occur.

More Information

CV   Paper