



The
Alan Turing
Institute



Opinions Are Not Always Positive: Debiasing Opinion Summarization With Model-Specific and Model-Agnostic Methods

Yanyue Zhang ^I, Yilong Lai ^I, Zhenglin Wang ^I, Pengfei Li ^I,
Deyu Zhou ^I, Yulan He ^{II}

^I School of Computer Science and Engineering, Key Laboratory of Computer Network and Information Integration, Ministry of Education, Southeast University, China

^{II} Department of Informatics, King's College London ^{II} The Alan Turing Institute, UK

Outline

1

Background

2

Methodology

3

Experiments

4

Conclusion

Outline

1

Background

2

Methodology

3

Experiments

4

Conclusion

Opinion Summarization

Reviews

Amazon Customer ★★★★★

Easy to put on and take off. Very protective glass against dust and scratches. Great price.

Amazon Customer ★★★★★

Fits like it's supposed to. It has definitely kept me from breaking my watch more than once. I seem to always run into something with my watch. Easy to pop on and off. Well pleased. Great price. As with Amazon and shipping now, took longer than 2 days. More like 4. Not bad but sure miss the 2 day shipping

Amazon Customer ★★★★★

This is a great cover for my Apple Watch. Easy to use and snap on. The screen is clear and the touch screen works perfectly. The best part is that it doesn't pop off like other ones. I would recommend to clean it every so often. Also it will break after extensive use because it is a snap on case.



summarizing user opinion from reviews
(e.g., a product, hotel, or restaurant).



with subjective emotions about an entity

Summary

Customers like the performance, fit, and ease of installation of the portable electronic device cover. For example, they mention it works well, it's secure, and it snaps on and off easily. Some are happy with value, and appearance. That said, opinions are mixed on quality, touch screen, and scratch resistance.

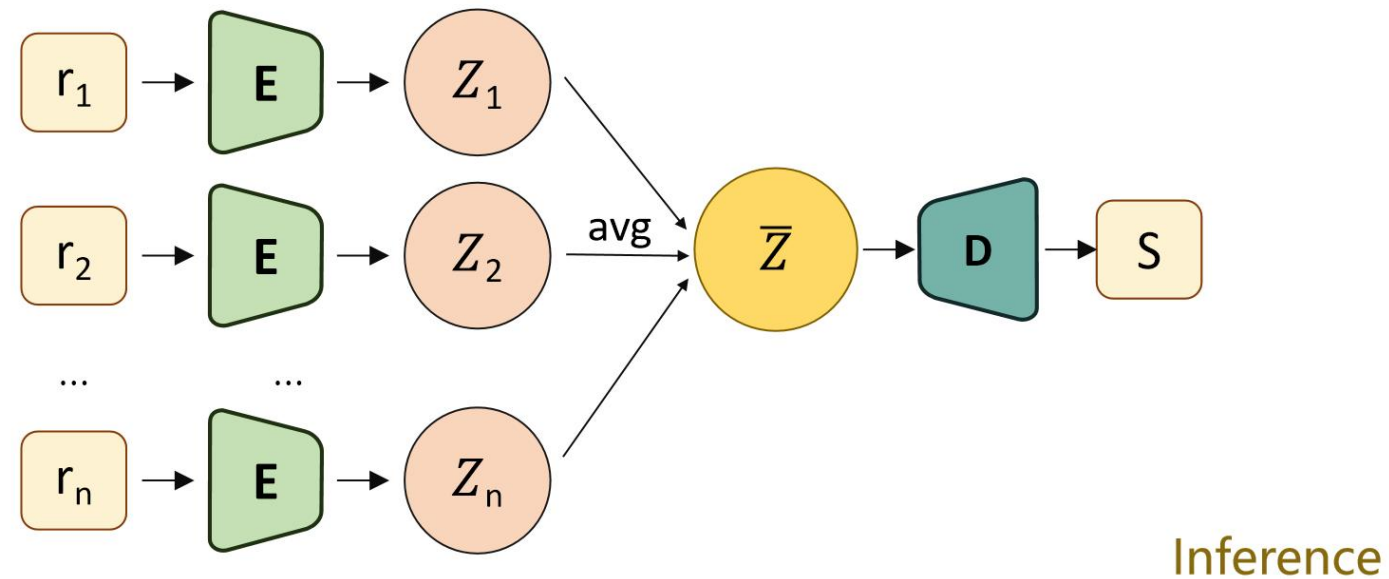
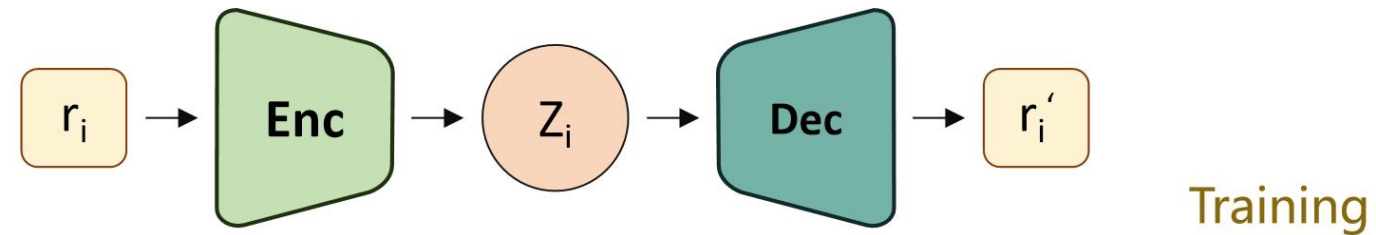
Hard to obtain

Previous Research

Architectures: Encoder-decoder based on AE\VAE\DAE

Training: Reconstruction Loss

Inference: Review aggregation



Emotional bias

Review ①	These are the bad tights for my 5-year old. The tights are badly made and can't last several washings (hang dry). The color is ugly , and my daughter hates ...
Review ②	These are my daughters common ballet tights. They can't fit well and squish her toes ...
Review ③	my 3 year old can't fit into these perfectly . I hate these tights , they are mismatched for wearing sandals to dance class ...
Review ④	Stiff fabric , runs small a though. U should consider getting one size up to avoid having your daughter, or son if he's into ballet, have little circulation. ...
Review ⑤	This is not my go to tight when my daughter needs new ones . I will buy anything else . They are super flimsy and last only one washing! ...
Coop (Iso et al., 2021)	These are great for the price. The tights are comfortable and don't take up much space. The only thing is that they can be worn to wear with the flip flops ... (I'm not sure if you have to wear them).
Trace (Zhang and Zhou, 2023)	These are great for those who want to wear a small. They are very comfortable and fit well. The only problem is that they don't last as long as some of the more expensive ones in the past . I would recommend these to anyone.

	Amazon		Yelp	
	Pos	Neg	Pos	Neg
Copycat	91.5	19	100	55.5
Coop	91.25	38.25	99.75	55.25
Wassos(T)	89.5	22.75	99.5	7.75
Wassos(O)	92.5	8	96.75	42.75
TRACE	92.5	31.75	99.75	56.5



The positive is too many !

83.5% in Amazon

72.3% in Yelp

Our Work

Existing bias mitigation methods can be broadly classified into two categories: model-specific and model-agnostic approaches.

Contribution:

- Discover the sentiment bias in opinion summarization.
- Propose the model-specific debias model DE-VAE based on sentiment disentanglement.
- Propose the model-agnostic debias methods PairDA based on counterfactual data augmentation via LLM.

Outline

1

Background

2

Methodology

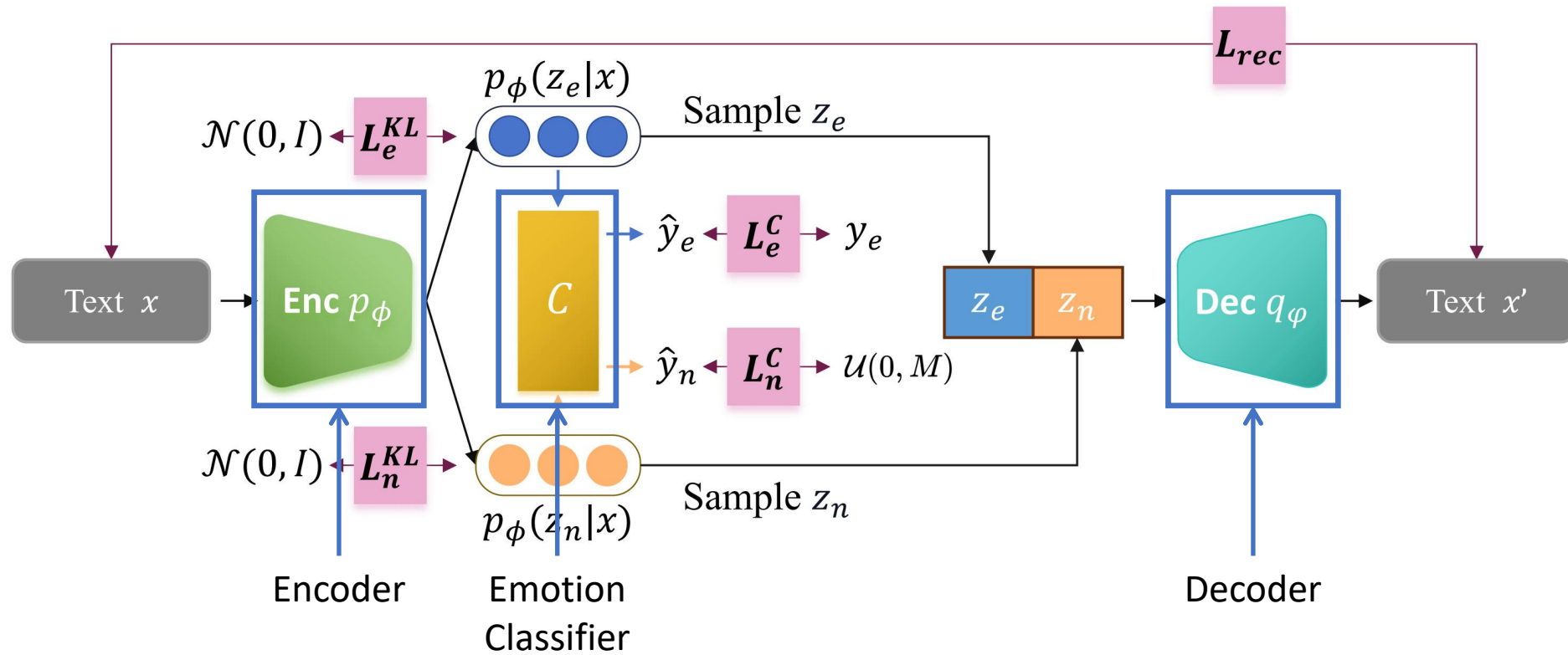
3

Experiments

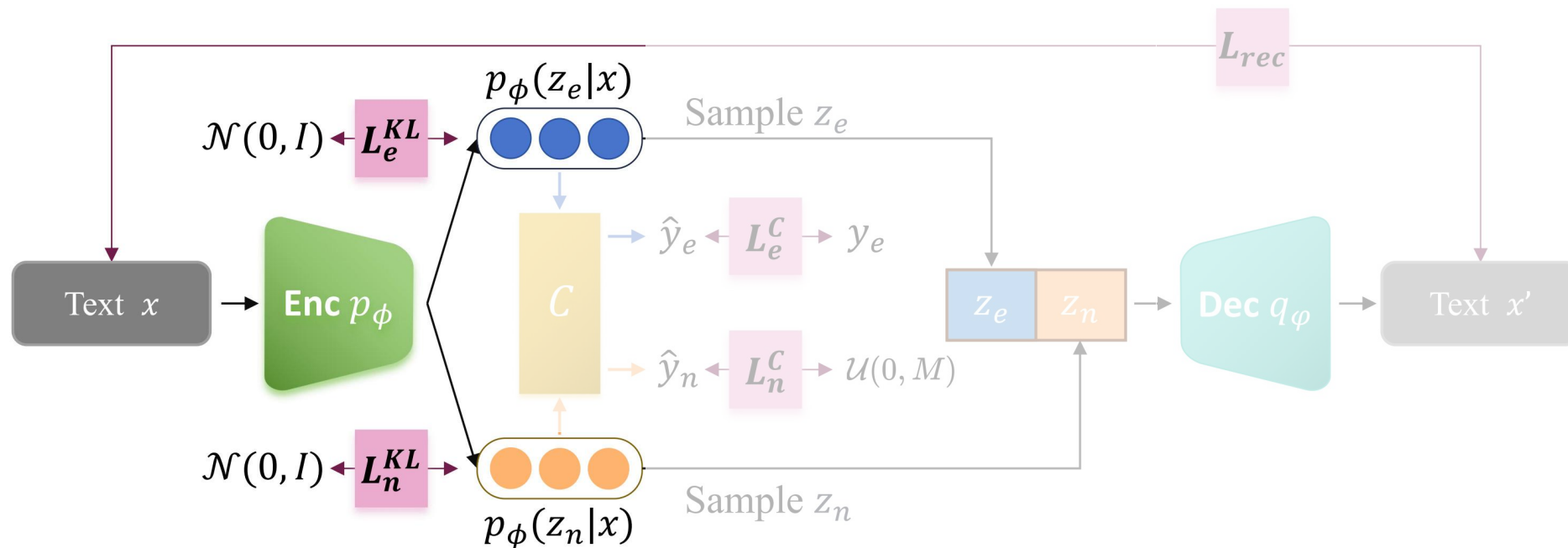
4

Conclusion

DE-VAE



DE-VAE

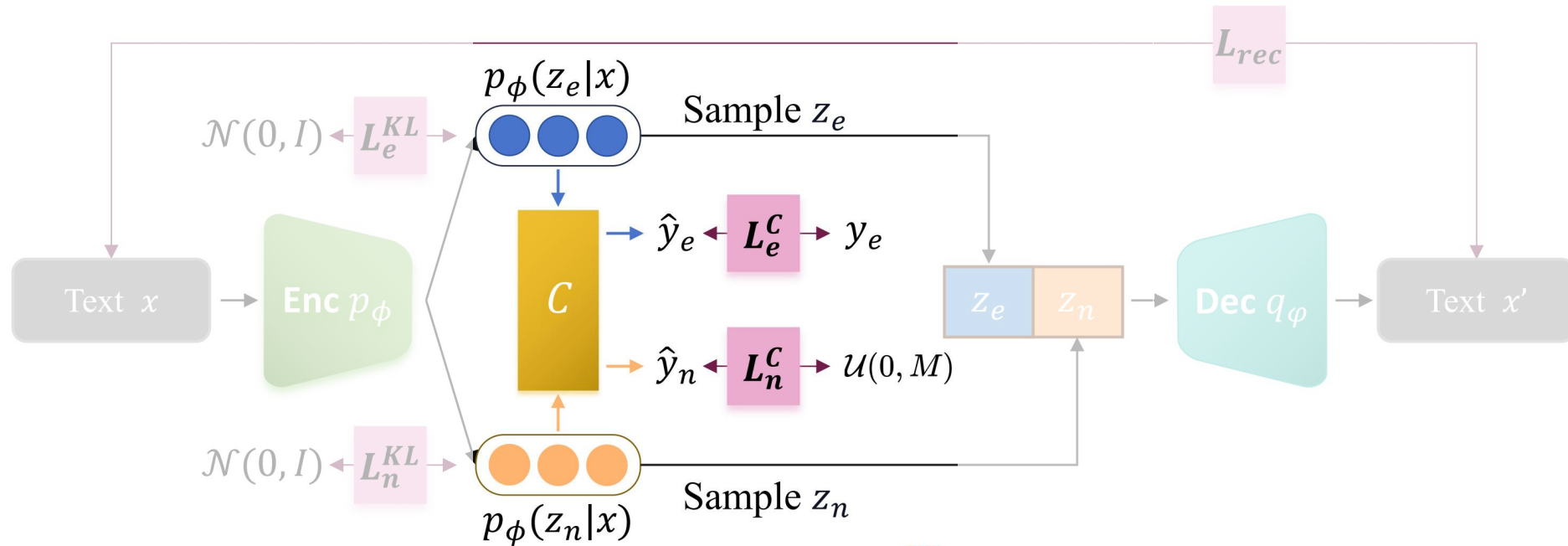


Kullback-Leibler loss based on the Kullback-Leibler (KL) divergence:

$$L_e^{KL} = \mathbb{D}_{KL}(p_\phi(z_e | x) || p(z_e))$$

$$L_n^{KL} = \mathbb{D}_{KL}(p_\phi(z_n | x) || p(z_n))$$

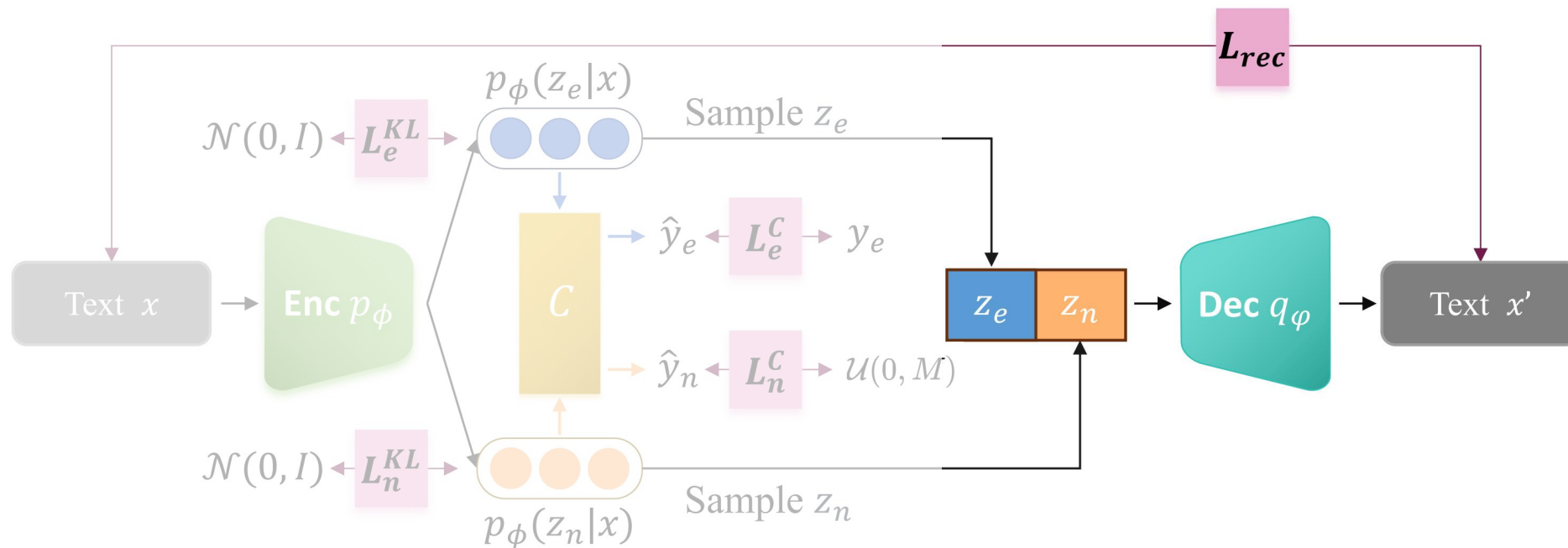
DE-VAE



Emotion relevant loss:
$$L_e^C = -\mathbb{E}_{p_\phi(z_e)} \sum_{i=1}^M y_{ci} \log(p(\hat{y}_e | z_e))$$

Emotion adversarial loss:
$$L_n^C = -\mathbb{E}_{p_\phi(z_n)} [\mathbb{D}_{KL}(\mathcal{U}(0, M) || p(\hat{y}_n | z_n))]$$

DE-VAE



Reconstruction loss:

$$L_{rec}(\phi, \varphi) = - \sum_{i=1}^N p_\phi(z|x) \mathbb{E} [\log q_\phi(x | z)]$$

PairDA

Foundational Prompt

Your task is to generate a counterfactual that retains internal coherence and avoids unnecessary changes.

Example: Really good movie. Maybe the best I've ever seen. Alien invasion, a la The Blob, with crazy good acting.....

Counterfactual: Really bad movie. Maybe the worst I've ever seen. Alien invasion, a la The Blob, without the acting.....

Example: This is one of the most awesome movies ever.....

.....

Text:Excellent and fresh ingredients, make this a must go to for tasty sushi.
Staff is unfriendly, but restaurant is spacious.

Counterfactual:

Learning The Difference That Makes A Difference With Counterfactually-Augmented Data

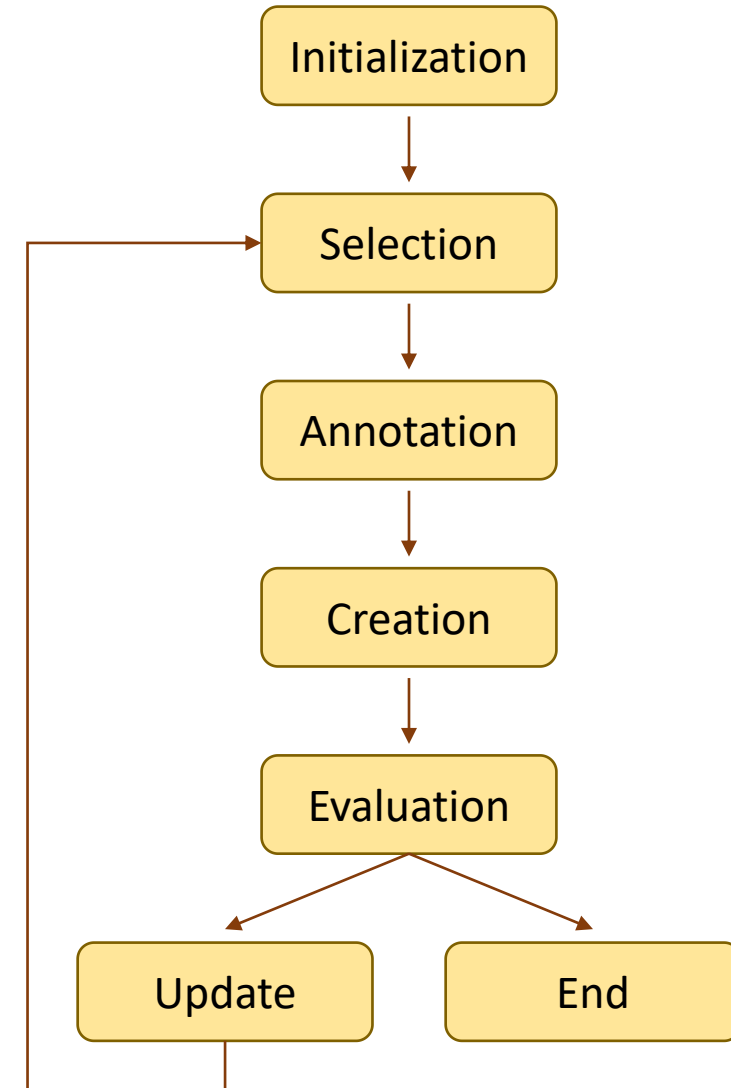


Mediocre and stale ingredients, make this a place to avoid for tasty sushi.
Although the staff is friendly, the restaurant is cramped.

Prompt Optimization

Step:

- (1) **Initialize** test set I , candidate set C and example sequence S .
- (2) **Select** a review x from C .
- (3) **Annotate** counterfactual y .
- (4) Insert (x, y) into S to **create** prompts.
- (5) **Evaluate** prompts based on I and manual assessment.
- (6) Based on the maximum score, decide whether to **stop**.
- (7) If not, **update** C and S based on x . Update I based on corresponding failed example from step (5). Then, to step (2).



Prompt Optimization

Algorithm 1 Prompt Optimization

Input: instruction D , test set $\mathcal{I} = \{x_1, \dots, x_{|\mathcal{I}|}\}$,
example permutation \mathcal{S} , candidate example set
 $\mathcal{C} = \mathcal{I}$, time step $t = 1$.

Output: Optimized Prompt $P \leftarrow P_t$.

- 1: **repeat**
- 2: randomly select review x_t from set \mathcal{C} and
obtained example $s(x_t, y_t)$ manually.
- 3: Insert $s(x_t, y_t)$ into \mathcal{S} to earned permutation
set $\{\mathcal{S}_t^1, \dots, \mathcal{S}_t^{|\mathcal{S}|+1}\}$, which each permuta-
tion contain $|\mathcal{S}| + 1$ examples.
- 4: **for** $i = 1$ to $|\mathcal{S}| + 1$ **do**
- 5: $P_t^i = \{D, \mathcal{S}_t^i\}$;
- 6: $score_t^i \leftarrow score(\{\mathcal{I} - \mathcal{S}\} | P_t^i)$;
- 7: **end for**
- 8: update permutation \mathcal{S} : $\mathcal{S} = \underset{\mathcal{S}_t^i}{argmax} score_t^i$;
- 9: $\mathcal{C} = \{\}$;
- 10: add x_i into \mathcal{C} if $score(x_i | P_t) < 0$;
- 11: $t = t + 1$;
- 12: **until** $score(\{\mathcal{I} - \mathcal{S}\} | P_t) > \delta$ or
 $score(\{\mathcal{I} - \mathcal{S}\} | P_t) - score(\{\mathcal{I} - \mathcal{S}\} | P_{t-1}) < \varepsilon$.

Based on the base prompt
and the current dataset,
ensuring containing samples
of varying generation difficulty.

Consider the position of
examples.

Evaluation based on
coherence and the success
rate of sentiment reversal.

Based on the current prompt
to reconstruct the test set,
reducing the number of tests
while requiring subsequent
additions to optimize the
current generation.

Outline

1

Background

2

Methodology

3

Experiments

4

Conclusion

Debias Performance

To conduct better sentiment testing, we have extracted additional sets of positive products (Pos) and negative products (Neg).

(%)	Amazon						Yelp					
	P	Pos R	F1	P	Neg R	F1	P	Pos R	F1	P	Neg R	F1
Copycat	91.5	53.04	67.16	19	69.09	29.80	100	69.20	81.80	55.5	100	71.38
Wassos(T)	89.5	53.67	67.10	22.75	68.42	34.15	99.5	51.89	68.21	7.75	93.94	14.32
Wassos(O)	<u>92.5</u>	50.14	65.03	8	51.61	13.85	96.75	62.82	76.18	42.75	92.93	58.56
TRACE	<u>92.5</u>	57.54	70.95	31.75	80.89	45.60	<u>99.75</u>	69.63	82.01	56.5	<u>99.56</u>	72.09
Coop(a)	84.75	55.48	67.06	32	67.72	43.46	100	53.48	69.69	13	100	23.01
Coop	91.25	59.64	72.13	38.25	81.38	52.04	99.5	68.74	81.31	54.75	99.10	70.53
PairDA	81.25	<u>82.28</u>	<u>81.76</u>	<u>82.5</u>	<u>81.48</u>	<u>81.99</u>	99.5	<u>93.21</u>	<u>96.25</u>	<u>92.75</u>	99.46	<u>95.99</u>
DE-VAE	95.25	98	96.61	98	98.25	98.12	100	98.50	99.24	98.5	98.25	98.38

Analysis

- Our PairDA and DE-VAE perform consistently well in almost all metrics, ranking first and second.
- DE-VAE exhibits better performance in sentiment debiasing compared to PiarDA.

Summarization Performance

We applied our pair wise counterfactual data augmentation method to enhance the reviews and summaries in the validation and test sets of both datasets.

	Amazon			Yelp		
	R1	R2	RL	R1	R2	RL
Copycat	31.7	6.0	20.3	26.0	5.2	18.2
Wassos(T)	29.5	6.3	19.9	31.1	5.6	18.5
Wassos(O)	31.5	6.9	21.0	26.2	4.3	16.1
TRACE	35.9	<u>7.1</u>	21.0	33.6	<u>6.6</u>	19.5
Coop(a)	32.9	6.0	20.8	31.6	6.2	<u>19.7</u>
Coop	<u>36.3</u>	7.0	<u>21.1</u>	<u>33.7</u>	6.4	19.5
PairDA	36.4	7.3	21.2	34.3	6.7	19.9
DE-VAE	34.2	6.7	21.0	33.1	6.2	19.0

Analysis

- PairDA outperforms PairDA notably in terms of ROUGE, and even brought about slight gains with the base model, Coop.

Data Augmentation Performance

(<i>%</i>)	<i>Reasonable</i>			<i>UnReasonable</i>		
	R_3	R_2	R_{total}	UR_3	UR_2	UR_{total}
<i>Amazon</i>	94.70	5.15	99.85	0.00	0.15	0.15
<i>Amazon_{DA}</i>	87.42	10.61	98.03	0.15	1.82	1.97
<i>Yelp</i>	95.72	3.67	99.39	0.00	0.61	0.61
<i>Yelp_{DA}</i>	91.78	7.44	99.22	0.11	0.67	0.78

	<i>Succ</i>	<i>Fail</i>
<i>Amazon</i>	80.80	19.20
<i>Yelp</i>	86.89	13.11

Analysis

- The data generated LLM has almost no coherence issues, but there is some discrepancy compared to the original data.
- There is still room for improvement in the success rate of sentiment transformation.

Outline

1

Background

2

Methodology

3

Experiments

4

Conclusion

Conclusion

Conclusion:

- Found noticeable sentiment bias in current opinion summarization models.
- Designed the Emotional Disentanglement VAE (DE-VAE).
- Introduced the method of counterfactual data augmentation through large models, PairDA.

Thank you