LASS: A Novel and Economical Data Augmentation Framework Based on Language Models for Debiasing Opinion Summarization

Yanyue Zhang[♠], Pengfei Li[♠], Yilong Lai[♠], Yulan He[♡] and Deyu Zhou[*♠]
[♠]School of Computer Science and Engineering, Key Laboratory of Computer Network and Information Integration, Ministry of Education, Southeast University, China
[♡]Department of Informatics, King's College London [♡]The Alan Turing Institute
{yanyuez98,lip.f,yilong.lai,d.zhou}@seu.edu.cn,
yulan.he@kcl.ac.uk

Abstract

As more than 70% of reviews in the existing opinion summary data set are positive, current opinion summarization approaches are hesitant to generate negative summaries given the input of negative texts. To address such sentiment bias, a direct approach without the reliance on a specific structure is to generate additional data based on large language models to balance the emotional distribution of the dataset. However, large-scale data augmentation based on large language models faces an apparent disadvantage, the expensive costs. Therefore, in this paper, we propose LASS, a novel data augmentation framework based on both **LA**rge and **S**mall language models for debia**S**ing opinion summarization. Specifically, a small number of synthesized negative reviews is obtained by rewriting the positive text via a large language model. Then, a disentangle reconstruction model is trained based on the generated data. After training, a large amount of synthetic data can be obtained by decoding the new representation obtained from the combination of different sample representations and filtering based on perplexity degree and sentiment classification. Experiments have proved that LASS can effectively alleviate emotional bias, similar to using only large models, but in a more economical way.

1 Introduction

With the unprecedented development of online interactive platforms, reviews on shopping platforms or social media become an important information source for manufacturers to make decisions. To cope with the flood of reviews, opinion summarization has received significant interest in natural language processing communities (Chu and Liu, 2019; Bražinskas et al., 2020; Amplayo and Lapata, 2020; Iso et al., 2021; Zhang and Zhou, 2023a). Unlike news, Wikipedia, and medical treatment records summarization, opinion summarization focuses on texts with user opinions and subjective emotions about an entity (e.g., a product, hotel, or restaurant). Accurately summarizing user perceptions and attitudes towards entities is a core requirement of opinion summarization.

	Amazon		Yelp	
	#Rev (K)	#Tok (M)	#Rev (K)	#Tok (M)
Data-Pos	1297	-	3367	-
Data-Neg	117	-	415	-
Bal Gen	1180	1049.2	2951	2669.7
Act Gen	540	397.6	630	564.6
LASS Use	200	178.1	200	179.5
Bal Gen	980	871.1	2751	2490.2
- LASS Use		**-83%**		**-93%**
Act Gen	340	299.1	430	385.1
- LASS Use		**-63%**		**-68%**

Table 1: Dataset and data augmentation analysis of the Amazon and Yelp dataset. '#Rev (K)' and '#Tok (M)' represent the number of reviews and the tokens processed by LLMs to generate the data, in thousands (K) and millions (M). 'Data-Pos' and 'Data-Neg' indicate the number of positive and negative samples in the dataset. 'Bal Gen', 'Act Gen', and 'Lass Use' represent the number of negative samples required to balance the dataset, for actual summarization training, and for LASS training, respectively.

However, as shown in previous work (Zhang et al., 2024), the current opinion summarization approaches are reluctant to generate a negative opinion summary given the input of negative opinions. To tackle such sentiment bias, Zhang et al. (2024) design a counterfactual data augmentation method via LLMs, PairDA, to directly alter the sentiment distribution of the dataset. However, data augmentation based on LLMs has a natural drawback, an expensive cost for large-scale data generation. As shown in Table 1, to achieve a fully balanced sentiment distribution dataset, millions of data points need to be generated (1.18M for Ama-

[*] Corresponding author.

Proceedings of the 31st International Conference on Computational Linguistics, pages 6169–6183
January 19–24, 2025.

zon and 2.95M for Yelp). While a full balance is not necessary to mitigate this issue, actual experience has shown that the amount of data needed to be generated still exceeds several million tokens, even reaching 564.6M on the Yelp dataset.

Therefore, in the paper, we propose LASS, a novel framework based on both **LA**rge and **S**mall language models for debia**S**ing opinion summarization. Firstly, a small size of synthesized negative reviews is obtained by rewriting the positive text via a large language model. Secondly, a disentangle reconstruction model is trained based on the generated data. Specifically, a disentangled autoencoder is proposed to obtain the sentiment and content representation through reconstruction, emotion, and distance constraints. Further, the new representations are obtained by exchanging the sentiment representation of the pair of counterfactual data, which are used to generate each other as the counterfactual reconstruction loss. To further constrain the emotion information, the original emotion representation is replaced with a learnable emotion label representation, where the weight depends on the outcome of emotion classification. Finally, a large amount of synthetic data can be obtained by decoding the new representation obtained from the combination of different sample representations and filtering based on perplexity and sentiment classification.

The experimental results demonstrate that LASS achieved results comparable to LLMs only, with an average reduction of 63.08% in synthetic data. Employing LASS for data augmentation across the four models resulted in an average increase of 33.7% in negative sentiment accuracy without affecting the Rouge scores (Lin, 2004) of the summaries, compared to 34.6% with LLMs only.

The main contributions of this paper are as follows:

- We propose LASS, a data augmentation framework combining large and small language models to alleviate emotional bias by optimizing the emotional distribution of datasets.
- We design a data reproduction method based on a disentangle reconstruction model, which generates additional data via decoding the combined new representations and filtering based on perplexity and sentiment classification.
- The experimental results demonstrate that LASS which combines large and small models can alleviate sentiment bias as effectively as the approach solely based on LLMs, but more economically.

2 Related Work

2.1 Opinion Summarization

Opinion summarization generally focuses on user reviews about products, hotels, restaurants, and so on. The abstractive approaches mainly utilize an encoder-decoder architecture, exploring various structures such as AE, VAE, or denoising autoencoder (DAE) (Chu and Liu, 2019; Bražinskas et al., 2020; Amplayo and Lapata, 2020; Iso et al., 2021; Zhang and Zhou, 2023a). During training, these models are constrained by the objective of reconstructing the input text, and during generation, they use the average of text representations as the summary representation for decoding. Subsequent approaches aimed to enhance the controllability of generating summaries by explicitly (Suhara et al., 2020; Elsahar et al., 2021; Amplayo et al., 2021a; Ke et al., 2022) or implicitly (Amplayo et al., 2021b) modeling aspect information. Some methods also explore ways to fuse input information for summarization beyond simple averaging, utilizing techniques like composite optimization (Iso et al., 2021), Wasserstein barycenter (Song et al., 2022), or hierarchical discrete latent space (Hosking et al., 2023). Recent work also has focused on the controllability and factuality of summaries (Zhang et al., 2024; Zhang and Zhou, 2023b; Hosking et al., 2023; Carichon et al., 2024; Syed et al., 2024; Benedetto et al., 2024).

2.2 Debiasing Strategies in NLP

Bias in NLP systems can typically be categorized as internal bias and external bias (Elsafoury et al., 2023; Li et al., 2023), depending on whether the bias is related to the training data of downstream tasks. Internal bias often pertains to issues of social fairness (Parraga et al., 2022), such as gender and racial bias, which have been identified in the embeddings of pre-trained language models (Guo et al., 2022). Existing work has attempted to address these issues through methods like adjusting pre-training data, introducing additional objectives, or post-processing (Li et al., 2023; Parraga et al., 2022; Elsafoury et al., 2023).

On the other hand, external bias related to downstream tasks is often associated with task-specific

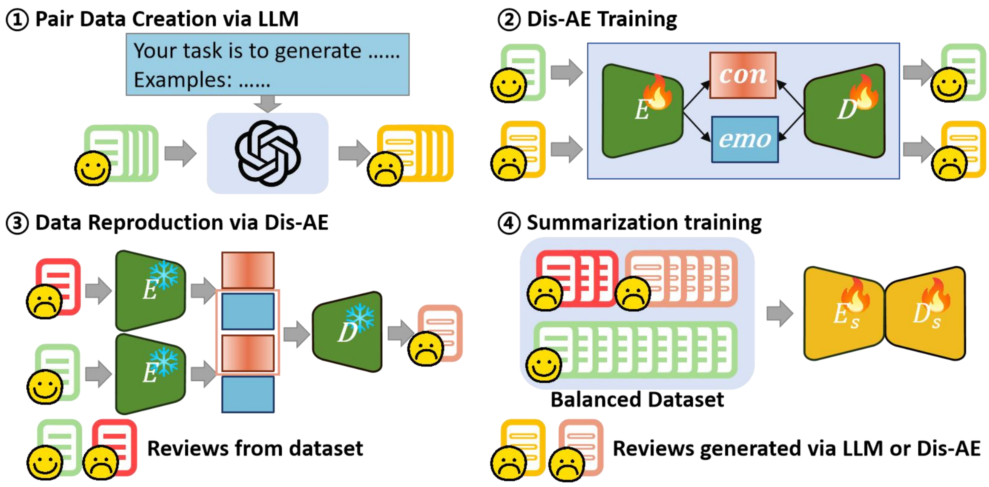


Figure 1: The pipeline of LASS. "E" and "D" represent the encoder and decoder, respectively. The Smiling and crying faces indicate the positive and negative emotional polarity of the text. The flame and snowflake symbols indicate whether the model is being trained or the parameters are frozen.

features, such as entity bias in fake news detection (Zhu et al., 2022), position bias in emotion cause extraction (Yan et al., 2021), and language bias in Visual Question Answering (VQA) (Cadene et al., 2019), and so on. To mitigate these specific biases, two distinct approaches have been developed: data distribution-related and model training-related (Shah et al., 2020; Parraga et al., 2022; Li et al., 2023). In the data distribution-related approach, efforts are made to re-sample, weight, or generate data to counteract bias (Dixon et al., 2018; Pruksachatkun et al., 2021; Qian et al., 2022). In contrast, model training-related methods explore adversarial techniques, causality (Cadene et al., 2019; Zhu et al., 2022), disentanglement, and additional auxiliary modules to mitigate bias.

3 Methodology

In this section, we describe LASS, the data augmentation debias method via both LLMs and a small generator, a disentangle autoencoder. As Figure 1 shows, the overall architecture of LASS contains four processes, pair data creation via LLMs, Dis-AE model training, data reproduction via Dis-AE, and summarization training. We first employ the LLMs with manual demonstrations to obtain pairs of counterfactual data in Section 3.1. In Section 3.2, based on pair samples, a disentanglement autoencoder, Dis-AE, is trained to obtain the sentiment and content representation through reconstruction, emotion, and distance constraints. After that, we introduce the data reproduction process via Dis-AE in Section 3.3 with the training. Finally, the generated data will be added to the original dataset to form an emotion-balanced dataset, which will be used to train any unsupervised summarization model.

3.1 Pair Data Creation via LLM

To avoid generating negative reviews that contain unreasonable product information, we obtain synthetic data by rewriting the original positive text following Zhang et al. (2024). Adhering to the principle of minimal modification, synthetic data with the opposite sentiment but identical content is generated through LLMs via prompt with manual demonstration.

3.1.1 Prompt Design

We first devised a foundational prompt to leverage the in-context learning capabilities of LLM for obtaining emotional opposite reviews. Then, we enhance the prompt design by incorporating human-annotated samples of corresponding datasets.

Formally, our foundational prompt is defined as a set $P = \{D, s(x_1, y_1), \cdots, s(x_k, y_k)\}$, comprising a task instruction D and k demonstration examples. $s(x_i, y_i)$ denotes a pairwise example of emotional counterfactuals. Specifically, we define task instruction D as *"Your task is to generate a counterfactual that retains internal coherence and avoids unnecessary changes."* and randomly select k samples from counterfactually-augmented movie reviews dataset (Kaushik et al., 2020), where $k = 5$.

Then, the current prompt is used to rewrite the reviews in the dataset, and a certain proportion of successfully and unsuccessfully rewritten reviews are randomly selected as a small evaluation dataset $\mathcal{I}$ for prompt enhancement. Starting with a new prompt $P = \{D\}$, counterfactual examples are generated by randomly selecting and annotating reviews from $\mathcal{I}$, which are then inserted to create the candidate prompts. The optimal prompt is selected based on the prompt's success rate in rewriting on the $\mathcal{I}$. The samples that failed to be rewritten are used as the new test set $\mathcal{I}$, and it is determined whether further annotation is needed. The more detailed steps of the procedure, prompts, and counterfactual examples are in appendix C, E, and E.1.

3.2 Dis-AE Training

Given a set of text pairs (user reviews) with the same content but opposite emotional polarity, Dis-AE aims to reconstruct the input pairs. As Figure 2 shows, the overall architecture of Dis-AE contains three components, an encoder p_θ, an emotional classifier C, and a decoder q_ϕ.

In the training stage, the positive text x^p is passed to the encoder $p_\theta(z_e, z_c \mid x)$ [1] to get two types of text representation, the sentiment z_e^p and the content z_c^p. Similarly, z_e^n and z_c^n can be obtained from x^n. Since the content of the paired texts is similar, but the emotion is opposite. Their content representations z_c^p and z_c^n are constrained to resemble each other, while their emotional representations z_e^p and z_e^n are forced to distance themselves.

All these representations are fed into the same emotional classifier C. To ensure that the emotion representation contains as little content information as possible, a learnable emotion label representation set Z_r is used to replace z_e^p and z_e^n. Z_r also constrains by emotion classification loss L_r and contains M emotion label representations, where M is the number of categories for emotion classification. M is the number of categories for emotion classification. Based on the emotion distribution $\hat{y}_e^p$ and $\hat{y}_e^n$ obtained by the corresponding emotion representation, the representation set Z^r is weighted to get the final emotion representation $\widetilde{z}_e^p$ and $\widetilde{z}_e^n$.

Then the document latent variable z^p is obtained by concatenating $\widetilde{z}_e^p$ and z_c^p, which is used to reconstruct the input text x^p through the decoder $q_\phi(x \mid z)$. Since pairs of text have similar content representations, combining another content representation z_c^e should also represent the current text x^p. Thus, positive counterfactual representations $\widetilde{z}^p$ are obtained by a combination of $\widetilde{z}_e^p$ and z_c^n, which is decoded to obtain x^p. Similarly, $\widetilde{z}_e^n$ is combined separately with z_c^n and z_c^p, and decoded to obtain x^c.

In order to ensure the basic ability of text generation, we retained the AE constraints, the reconstruction loss L_{rec}. When reconstructing the input pair separately, representation z^p from concatenated z_e^p and z_c^p is used as the input of the decoder to reconstruct the input text x^p. The same procedure is applied to obtain the corresponding negative text x^n. The reconstruction loss is defined as:

$$
\begin{aligned}
&L_{rec}(\theta, \phi) = \\
&- \sum_{i=1}^{N} \mathop{\mathbb{E}}_{p_\theta(\widetilde{z}_e^p, z_c^p \mid x^p)} [\log q_\phi(x^p \mid \widetilde{z}_e^p, z_c^p)] \\
&- \sum_{i=1}^{N} \mathop{\mathbb{E}}_{p_\theta(\widetilde{z}_e^n, z_c^n \mid x^n)} [\log q_\phi(x^n \mid \widetilde{z}_e^n, z_c^n)],
\end{aligned}
\tag{1}
$$

where θ and ϕ are the parameters of the model. The reconstruction loss improves the quality of the decoded text and forces the text representation to store content information with emotion. To disentangle emotional representation and content representation, we employ an emotional auxiliary constrain $\mathcal{L}_{emo}$, which is including with emotion classification constraints L_e, emotion adversarial constraints L_c and label emotion constraints L_r.

The sentiment representation z_e^p and z_e^n and content representation z_c^p and z_c^n are fed into classifier C separately. The prediction result of z_e^p and z_e^n should be the corresponding emotion label y_e and y_n, which is a cross-entropy loss:

$$
\begin{aligned}
L_e(\theta) = &- \mathbb{E}_{p_\theta(z_e^p)} \sum_{i=1}^{M} y_e^p log(p(\hat{y}_e^p | z_e^p)) \\
&- \mathbb{E}_{p_\theta(z_e^n)} \sum_{i=1}^{M} y_e^n log(p(\hat{y}_e^n | z_e^n)).
\end{aligned}
\tag{2}
$$

Inspired by Pergola et al. (2021), rather than being unable to achieve correct classification, we assume that content representations z_c^p and z_c^n are sentiment-neutral, and should not exhibit any category bias during sentiment classification. Therefore, z_c^p and z_c^n should be fed into the sentiment

[1] x represents x^p or x^n. Similarly, z_e and z_c represent z_e^p, z_e^n, z_c^p, and z_c^n, respectively. z represents z^p or z^n.

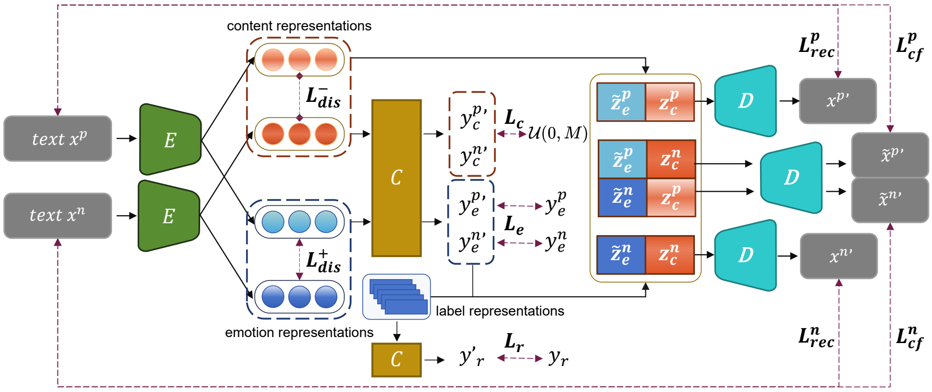


Figure 2: The architecture of the disentanglement Model, Dis-AE. E and D are the encoder and the decoder. y_e^p, y_e^n and y_r are the emotion labels corresponding to the input x^p, x^n and label representation. C is a sentiment classifier. M is the number of categories for emotion classification.

classifier to obtain a uniform sentiment classification distribution, which is an expected KL divergence loss:

$$L_n(\theta) = -\mathbb{E}_{p_\theta(z_c^p)}[\mathbb{D}_{KL}(\mathcal{U}(0, M)||p(\hat{y}_c^p|z_c^p))] - \mathbb{E}_{p_\theta(z_c^n)}[\mathbb{D}_{KL}(\mathcal{U}(0, M)||p(\hat{y}_c^n|z_c^n))], \tag{3}$$

where M is the total number of sentiment classes. The former is the expected KL divergence with the uniform distribution $\mathcal{U}(0, M)$. Given that an additional learnable label representation set $Z^r = \{z_1^r, \cdots, z_M^r\}$ is used to replace the emotion representations z_e^p and z_e^n, Z^r also need to contain emotional information constrained by a similar loss of emotional classification:

$$L_r = -\sum_{i=1}^{M} y_i^r log(p(\hat{y}_i^r|z_i^r)). \tag{4}$$

To further introduce relational knowledge hidden in pairs of data, we add distance loss $\mathcal{L}_{dis}$ and counterfactual reconstruction loss $\mathcal{L}_{cf}$. The distance loss is based on the prior knowledge that the input text pair expresses opposite emotions but shares similar content. The represented distance is constrained based on the sentence similarity:

$$L_{dis} = 2 + sim(z_e^p, z_e^n) - sim(z_c^p, z_c^n), \tag{5}$$

where $sim(\cdot)$ indicates the cosine similarity function. Likewise, since the text pair x^p and x^n contain the same content information, the alternate content representation should allow for successful decoding of the corresponding text. Thus the counterfactual reconstruction loss is:

$$L_{cf}(\theta, \phi) = -\sum_{i=1}^{N} \mathop{\mathbb{E}}_{p_\theta(\tilde{z}_e^p, z_c^p|x^p) p_\theta(\tilde{z}_e^n, z_c^n|x^n)} [\log q_\phi(x^p \mid \tilde{z}_e^p, z_c^n)] - \sum_{i=1}^{N} \mathop{\mathbb{E}}_{p_\theta(\tilde{z}_e^n, z_c^n|x^n) p_\theta(\tilde{z}_e^p, z_c^p|x^p)} [\log q_\phi(x^n \mid \tilde{z}_e^n, z_c^p)]. \tag{6}$$

Our final objective function is:

$$\mathcal{L} = L_{rec} + \alpha\mathcal{L}_{emo} + \beta\mathcal{L}_{dis} + \gamma\mathcal{L}_{cf}, \tag{7}$$

where α, β and γ are hyper-parameters that controls the strength of constrains.

3.3 Data Reproduction via Dis-AE

After training, data reproduction can be performed by selecting parent samples from the training set and combining them with the disentanglement model Dis-AE. Specifically, when negative reviews for a specific product are needed, positive reviews for that product are selected along with any negative reviews as parents. The parent samples are inputted into Dis-AE to obtain sentiment representations and content representations separately. By combining the content representation of positive reviews with the sentiment representation of negative reviews, we obtain the child representation. Decoding the child representation yields negative samples. This data reproduction approach ensures the controllability of content and sentiment of generated

	Amazon						Yelp					
		Pos			Neg			Pos			Neg	
(%)	Rev	Sen	Dif	Rev	Sen	Dif	Rev	Sen	Dif	Rev	Sen	Dif
Wassos(T)	93.25	88.97	-	20.63	19.84	-	98.25	91.51	-	43.5	47.25	-
TRACE(a)	91.63	82.29	-	24.38	29.61	-	100	94.53	-	68.5	57.08	-
TRACE	89.25	80.94	-	40.5	38.82	-	99.5	97.44	-	8.5	10.92	-
Wassos(O)	**93.5**	**92.49**	-	7.13	10.31	-	79.25	78.93	-	59.25	53.28	-
+GPT	62.50	64.18	**-27.8**	69.38	**49.78**	**+38.3**	**90.63**	76.31	+4.4	68.38	59.33	+7.6
+LASS	49.75	63.69	−34.4	**69.75**	47.06	+38.2	89.38	**80.04**	**+5.6**	**83.63**	**71.83**	**+21.5**
Copycat	**93.75**	**84.69**	-	16.25	16.40	-	**97.75**	**88.43**	-	47.75	41.15	-
+GPT	60.95	57.00	−30.3	70.63	55.09	+46.5	95.00	76.71	**-7.2**	78.13	63.96	+26.6
+LASS	61.13	64.01	**-26.7**	**76.75**	**58.34**	**+51.2**	93.38	76	−8.4	**86.50**	**64.02**	**+30.8**
Coop(a)	81.75	76.05	-	46.88	41.39	-	99.88	93.23	-	34	39.31	-
+GPT	**94.38**	**88.26**	**+12.4**	**90.88**	**79.68**	**+41.1**	99.63	94.98	+0.7	77.50	**73.48**	+38.8
+LASS	92.75	86.42	+10.7	89.38	74.14	+37.6	**99.75**	**97.26**	**+2.0**	**79.25**	72.37	**+39.2**
Coop	82.75	76.97	-	58	47.64	-	99	92.38	-	51.5	47.55	-
+GPT	90.63	81.40	+6.2	**93**	**76.48**	**+31.9**	**100**	**95.52**	**+2.1**	**93.38**	**80.86**	**+37.6**
+LASS	**90.75**	**82.38**	**+6.7**	84.88	68.89	+24.1	99.75	95.41	+1.9	90.25	77.13	+34.2

Table 2: Sentiment accuracy results on Amazon and Yelp. The bold scores denote the best scores.

text while also meeting the demand for large-scale data augmentation, due to the diversity of parental sample combinations.

Due to the limitation of small model generation ability, the generated text may be unreadable, or with incorrect sentiment polarity. Therefore, we add a data filtering process based on perplexity and sentiment classification to ensure the quality of the generated text.

4 Experiments

4.1 Datasets

We performed experiments on two opinion summarization benchmarks, the Amazon dataset (Bražinskas et al., 2020) and Yelp (Chu and Liu, 2019). All datasets include review ratings with a 1–5 scale which we used as sentiment labels. Besides training reviews, these two datasets also contain gold-standard summaries for 200 and 60 sampled objects for evaluation. Additionally, we use two specialized review sets, the positive (POS) and the negative (NEG), to evaluate the summarization model's ability to generate positive and negative summaries. Following Zhang et al. (2024), the sets include 800 positive or 800 negative products from the training data. Half are for validation, and the other half for testing. Each product consists of 7 or 8 reviews, all rated as 5 for positive or 1 for negative sentiment. Because of the consistent sentiment polarity of reviews, they are able to assess the ability of the summarization model to generate summaries with different sentiments.

	Amazon			Yelp		
	R1	**R2**	**RL**	**R1**	**R2**	**RL**
Wassos(T)	29.7	6.5	20.0	30.8	5.9	18.3
TRACE(a)	33.7	6.3	20.5	32.6	6.6	20.0
TRACE	36.0	7.2	20.8	33.9	6.8	19.7
Wassos(O)	31.5	**6.9**	**21.0**	26.6	4.5	16.4
+GPT	32.4	6.1	19.9	27.3	5.3	18.8
+LASS	**32.7**	6.1	19.5	**29.7**	**6.1**	**19.0**
Copycat	31.9	6.1	**20.4**	29.3	5.4	17.7
+GPT	32.3	5.9	19.7	**30.0**	5.6	18.8
+LASS	**32.6**	**6.4**	19.8	29.4	**6.0**	**19.2**
Coop(a)	32.1	5.1	18.1	30.6	5.9	18.8
+GPT	32.2	**7.1**	20.2	**31.6**	6.4	19.5
+LASS	**32.9**	6.3	**20.4**	31.6	**6.9**	**19.6**
Coop	35.7	6.2	19.8	**34.5**	**6.9**	**19.6**
+GPT	35.6	6.4	20.6	34.0	6.8	19.5
+LASS	**36.2**	**7.0**	**21.4**	33.8	6.9	19.4

Table 3: Rouge scores on Amazon and Yelp. The bold scores denote the best scores.

4.2 Evaluation Metrics and Baselines

We evaluate summary systems with the classical ROUGE-1, 2, L metrics (Lin, 2004). We also report sentiment precision about the positive and the negative at the sentence level (Sen) and review level (Rev), using the sentiment analysis model from Stanza (Qi et al., 2020) to compute. All ratings are normalized to scores between 0 and 1. More details are in the Appendix A. The term "Dif" represents the average change in sentiment accuracy at both the review and sentence levels after data augmentation using GPT or LASS.

	Amazon		Yelp	
	#Rev (K)	#Tok (M)	#Rev (K)	#Tok (M)
Act Gen				
copycat	540	477.2	630	564.6
coop	450	397.6	630	564.6
LASS Use	200	178.1	200	179.5
Act Gen - LASS Use				
copycat	340	299.1	430	385.1
	-63.0%	**-62.7%**	**-68.2%**	**-68.2%**
coop	250	219.6	430	385.1
	-55.6%	**-55.2%**	**-68.2%**	**-68.2%**

Table 4: Data augmentation analysis of Amazon and Yelp dataset. '#Rev (K)' and '#Tok (M)' represent the number of reviews and the tokens processed by LLMs to generate the data, in thousands (K) and millions (M). 'Act Gen', and 'Lass Use' represent the number of negative samples used for the actual training process of coop and copycat, and used during Dis-AE training, respectively.

Following prior work (Iso et al., 2021; Song et al., 2022), we compare with **Copycat** (Bražinskas et al., 2020), **Coop** (Iso et al., 2021), **Wassos** (Song et al., 2022) and TRACE (Zhang and Zhou, 2023a). (a), (O), and (T) represent different clustering strategies for the model. The detailed introduction is in the Appendix B. Considering the sensitivity of the counter-templates in TRACE to training data, we experimented with data augmentation methods based on ChatGPT and LASS on three models Coop, Coop(a), and Copycat.

4.3 Implementation Details

In this work, we employ the ChatGPT platform [2] to generate pairwise emotional counterfactuals within a crafted prompt setting. Furthermore, the temperature parameter is designated as $T = 0.2$ to encourage a more deterministic output from the language model. For prompt optimization, the parameters are set as $m = 40$, $n = 10$, $\delta = 80\%$ and $\varepsilon = 10\%$. The final prompts include 5 pairs of examples for the Amazon dataset and 7 pairs for Yelp. Specifically, we extract the samples with a sentiment score of 5 from the training data.

For the disentanglement model Dis-AE, we used Adam optimizer (Kingma and Ba, 2015) with a linear scheduler, whose initial learning rate is set to $5e^{-4}$. For beam search in the generation, the beam size is set to 4 and a max token size of 70. The amount of training data used is 200k, according to the analysis in Section 5. Additionally, based on the PPL testing conducted on the training set, we set the threshold for PPL at 125. Only generated samples with PPL less than 125 and classified as negative by review level sentiment classifier from Stanza (Qi et al., 2020) are retained. To prevent the imbalance of multiple constraints from undermining the text generation capability, we mimic KL annealing (Li et al., 2019; Iso et al., 2021) to gradually increase α, β, and γ from 0 during training. The upper limit for the weight of sentiment loss α is set to 5, while β and γ are both limited to 1. All experiments were conducted on NVIDIA GeForce RTX 3090 or NVIDIA Tesla V100.

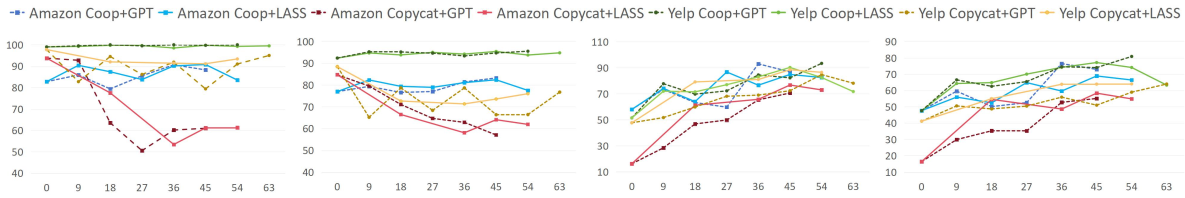
[2] https://chat.openai.com/chat

4.4 Results

According to Table 2, synthesized data from both LASS and GPT significantly enhance the model's performance in nearly all accuracy measures, whether at the review or sentence level. Specifically, the use of LASS achieved comparable results to GPT. However, for positive accuracy, the data augmentation methods had a negative impact on some models. For the Wassos(O) and Copycat models on the Amazon dataset, the data augmentation methods, whatever GPT or LASS, reduced the positive sentiment accuracy to around 60%, and the improvement in negative sentiment accuracy failed to reach 80%.

This kind of exception may be attributed to the multiple influences of data augmentation methods, summarization models, and datasets. From the perspective of summarization models, the overall performance of the Wassos(O) and the Copycat are inferior to that of Coop(a) and Coop in terms of both sentiment accuracy and ROUGE scores. For positive sentiment accuracy, any model's performance on the Yelp dataset as a whole is better than that on Amazon. This may be because the Yelp data mainly consists of restaurant reviews, making it easier for models to learn expressions of positivity and negativity compared to the diverse product types in the Amazon data.

From the ROUGE scores in Table 3, it was observed that all methods did not exhibit a performance decrease after data augmentation using GPT or LASS. This suggests that the data augmentation methods are applicable across different models and do not degrade the performance of the models on the original task. It also indicates that both the GPT and LASS methods generate highly readable data, and even with a large-scale addition to the training data, they do not disrupt the training of



(a) Positive-Rev (b) Positive-Sen (c) Negative-Rev (d) Negative-Sen

Figure 3: Experimental results about Coop and Copycat with different sizes of synthesized data from GPT or LASS on Amazon and Yelp. The horizontal axis represents the amount of synthetic data added, with each unit representing 10k (e.g., 9 represents 90k).

$Num(k)$	**PPL** ↓	**R1**	**R2**	**RL**
50	540.25	53.61	16.80	34.74
100	1314.50	60.57	23.31	42.65
150	788.96	56.25	34.08	50.92
200	360.13	59.55	38.85	54.86
250	403.57	59.29	39.36	54.44

Table 5: Experimental results about Dis-AE with different sizes of train data on Amazon.

summarization tasks.

5 Analysis

5.1 The Impact of LASS on Data Generation Requirements

To evaluate the impact of using LASS on the data generation requirements for LLMs, we analyzed the amount of data that large models need to generate with or without LASS. During the training of each summarization model, we tested the actual amount of synthetic data required by gradually adding more augmented negative reviews (Act Gen). By calculating the difference between the 'Act Gen' and the data used by LASS, we can determine the proportion of synthetic data saved by using LASS. To provide a clearer comparison of the LLM's resource consumption, we calculated the total number of tokens consumed when generating counterfactual text for each review (#ToK), including the input review text, the prompt text, and the output counterfactual text. The complete table can be found in the appendix D.

As shown in Table 4, we found that the amount of data required to address bias in the dataset is significantly less than the amount needed to balance the positive and negative data at a 1:1 ratio. For example, to fully balance the sentiment distribution in the Amazon dataset, 1.18 million synthetic reviews would be required. However, during the actual training of Coop and Copycat, once 0.54M and 0.45M synthetic data points were added to the original dataset, there was no further improvement in sentiment accuracy. Using LASS, the number of tokens required was reduced by 55%–68%, significantly reducing the demand for large models to generate synthetic data.

5.2 Effect of Augmented Data Volume on Summarization

To investigate the impact of synthetic data on summarization models, we analyzed the sentiment accuracies of different summarization models using varying amounts of augmentation data from GPT or LASS, as shown in Figure 3. Overall, adding negative reviews can improve the accuracy of negative sentiment of summaries, but may somewhat affect the ability to generate positive summaries to some extent. For Coop, the positive accuracy on Amazon shows some instability as the data volume increases. Meanwhile, Copycat's positive accuracy experiences a significant decline, suggesting that Copycat may not handle sentiment information well in summaries and tends to generate neutral text with mixed positive and negative sentiments.

5.3 Effect of Augmented Data Volume on Dis-AE.

Additionally, we explored the amount of data required for training Dis-AE. Evaluating whether the quality of the generated text meets the training requirements of summarization requires a lot of downstream experiments. To more efficiently confirm the data requirements, we employ two metrics: perplexity (PPL) and counterfactual reconstruction ROUGE score. The counterfactual reconstruction ROUGE score is similar to the counterfactual reconstruction loss L_{cf}, calculating the ROUGE score of reconstructed text after exchanging paired counterfactual samples with target text. PPL relies on GPT-2 to compute the degree of text fluency [3].

[3]https://huggingface.co/docs/transformers/perplexity

The results, as shown in Table 5, indicate that the quality of generation improves steadily with the increase in data volume, with instabilities observed after reaching 200k. The reason why the PPL for 50k is less than that for 100k is because samples shorter than 10 characters are not included in the PPL calculation, as PPL becomes erratic for excessively short texts.

6 Limitation

Overall, while debias through data augmentation can generalize across different models, its effectiveness is also limited by the performance and characteristics of each model. For example, in the current scenario, the Copycat model experienced significant degradation in positive sentiment accuracy after using augmented data on the Amazon dataset. For another model TRACE, changes in data distribution significantly affect the performance of the summaries, as observed in our preliminary experiments. This may be attributed to one of the parameters, the counter-template, being sensitive to the training data. Additionally, determining the minimum data required for Dis-AE training is a critical issue. The current approach, based on perplexity and counterfactual reconstruction metrics, only indirectly reflects the quality of generated counterfactual texts. We will continue to explore the training data requirement for Dis-AE in future work.

7 Conclusion

We propose LASS, a data augmentation framework that combines large and small language models to alleviate emotional bias by optimizing the emotional distribution of datasets. Leveraging a disentanglement reconstruction model, we design a novel data augmentation method, which generates additional data via decoding the combined new representations and filtering based on perplexity and sentiment classification. Experimental results show that LASS can alleviate sentiment bias as effectively as the approach solely based on LLMs, but more economically. By using a small disentangled model for data reproduction, LASS demonstrates that small models can achieve the same capabilities as LLMs through architecture and method design.

8 Acknowledgements

We would like to thank anonymous reviewers for their valuable comments and helpful suggestions. The authors acknowledge financial support from the National Natural Science Foundation of China (62176053). This research work is also supported by the Big Data Computing Center of Southeast University. YH was supported by a Turing AI Fellowship (EP/V020579/1, EP/V020579/2) funded by the UK Research and Innovation.

References

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021a. Aspect-controllable opinion summarization. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021b. Unsupervised opinion summarization with content planning. In Proceedings of the AAAI Conference on Artificial Intelligence.

Reinald Kim Amplayo and Mirella Lapata. 2020. Unsupervised opinion summarization with noising and denoising. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

Irene Benedetto, Moreno La Quatra, Luca Cagliero, Luca Vassio, and Martino Trevisan. 2024. Tasp: Topic-based abstractive summarization of facebook text posts. Expert Systems with Applications, 255:124567.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Unsupervised opinion summarization as copycat-review generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. 2019. Rubi: Reducing unimodal biases for visual question answering. Advances in neural information processing systems, 32.

Florian Carichon, Chrys Ngouma, Bang Liu, and Gilles Caporossi. 2024. Objective and neutral summarization of customer reviews. Expert Systems with Applications, 255:124449.

Eric Chu and Peter Liu. 2019. Meansum: A neural model for unsupervised multi-document abstractive summarization. In International Conference on Machine Learning, pages 1223–1232. PMLR.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pages 67–73.

Fatma Elsafoury, Stamos Katsigiannis, and Naeem Ramzan. 2023. On bias and fairness in nlp: How to have a fairer text classification? arXiv e-prints, pages arXiv–2305.

Hady Elsahar, Maximin Coavoux, Jos Rozen, and Matthias Gallé. 2021. Self-supervised and controlled multi-document opinion summarization. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Auto-debias: Debiasing masked language models with automated biased prompts. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1012–1023.

Tom Hosking, Hao Tang, and Mirella Lapata. 2023. Attributable and scalable opinion summarization. arXiv preprint arXiv:2305.11603.

Hayate Iso, Xiaolan Wang, Yoshihiko Suhara, Stefanos Angelidis, and Wang-Chiew Tan. 2021. Convex aggregation for opinion summarization. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 3885–3903.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In International Conference on Learning Representations.

Wenjun Ke, Jinhua Gao, Huawei Shen, and Xueqi Cheng. 2022. Consistsum: Unsupervised opinion summarization with the consistency of aspect, sentiment and semantic. In Proceedings of the fifteenth ACM international conference on web search and data mining, pages 467–475.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In ICLR (Poster).

Bohan Li, Junxian He, Graham Neubig, Taylor Berg-Kirkpatrick, and Yiming Yang. 2019. A surprisingly effective fix for deep latent variable modeling of text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3603–3614.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2023. A survey on fairness in large language models. arXiv preprint arXiv:2308.10149.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, pages 74–81.

Otávio Parraga, Martin D More, Christian M Oliveira, Nathan S Gavenski, Lucas S Kupssinskü, Adilson Medronha, Luis V Moura, Gabriel S Simões, and Rodrigo C Barros. 2022. Debiasing methods for fairer neural models in vision and language research: A survey. arXiv preprint arXiv:2211.05617.

Gabriele Pergola, Lin Gui, and Yulan He. 2021. A disentangled adversarial neural topic model for separating opinions from plots in user reviews. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2870–2883.

Yada Pruksachatkun, Satyapriya Krishna, Jwala Dhamala, Rahul Gupta, and Kai-Wei Chang. 2021. Does robustness improve fairness? approaching fairness with word substitution robustness methods for text classification. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3320–3331.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Michael Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer nlp. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9496–9521.

Deven Santosh Shah, H Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5248–5264.

Jiayu Song, Iman Munire Bilal, Adam Tsakalidis, Rob Procter, and Maria Liakata. 2022. Unsupervised opinion summarisation in the wasserstein space. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 8592–8607.

Yoshihiko Suhara, Xiaolan Wang, Stefanos Angelidis, and Wang-Chiew Tan. 2020. Opiniondigest: A simple framework for opinion summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

Ayesha Ayub Syed, Ford Lumban Gaol, Alfred Boediman, and Widodo Budiharto. 2024. Airline reviews processing: Abstractive summarization and rating-based sentiment classification using deep transfer learning. International Journal of Information Management Data Insights, 4:100238.

Hanqi Yan, Lin Gui, Gabriele Pergola, and Yulan He. 2021. Position bias mitigation: A knowledge-aware graph model for emotion cause extraction. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3364–3375.

Yanyue Zhang, Yilong Lai, Zhenglin Wang, Pengfei Li, Deyu Zhou, and Yulan He. 2024. Opinions are not always positive: Debiasing opinion summarization with model-specific and model-agnostic methods. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 12496–12513, Torino, Italia. ELRA and ICCL.

Yanyue Zhang and Deyu Zhou. 2023a. Disentangling text representation with counter-template for unsupervised opinion summarization. In Findings of the Association for Computational Linguistics: ACL 2023, pages 6344–6357.

Yanyue Zhang and Deyu Zhou. 2023b. Disentangling text representation with counter-template for unsupervised opinion summarization. In Findings of the Association for Computational Linguistics: ACL 2023, pages 6344–6357, Toronto, Canada. Association for Computational Linguistics.

Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. 2022. Generalizing to the future: Mitigating entity bias in fake news detection. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 2120–2125.

Algorithm 1 Prompt Optimization

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

Ensure: Optimized Prompt $P \leftarrow P_t$.

repeat
- randomly select review x_t from set $\mathcal{C}$ and obtained example $s(x_t, y_t)$ manualy.
- Insert $s(x_t, y_t)$ into $\mathcal{S}$ to earned permutation set $\{\mathcal{S}_t^1, \cdots, \mathcal{S}_t^{|s|+1}\}$, which each permutation contain $|\mathcal{S}| + 1$ examples.
- **for** $i = 1$ to $|\mathcal{S}| + 1$ **do**
  - $P_t^i = \{D, \mathcal{S}_t^i\}$;
  - $score_t^i \leftarrow score(\{\mathcal{I} - \mathcal{S}\}|P_t^i)$;
- **end for**
- update permutation $\mathcal{S}$: $\mathcal{S} = \underset{\mathcal{S}_t^i}{argmax}\ score_t^i$;
- $\mathcal{C} = \{\}$;
- add x_i into $\mathcal{C}$ if $score(x_i|P_t) < 0$;
- $t = t + 1$;

until $score(\{\mathcal{I} - \mathcal{S}\}|P_t) > \delta$ or $score(\{\mathcal{I} - \mathcal{S}\}|P_t) - score(\{\mathcal{I} - \mathcal{S}\}|P_{t-1}) < \varepsilon$.

A Sentiment Evaluation

For positive reviews, the sentiment score is 1 for positive, 0.5 for neutral, and 0 for negative, while for the negative set, the negative is 1. The rating for the review level precision involves assigning a score to the entire text, while at the sentence level, scores are assigned to each sentence in the text and then averaged.

B Baselines

We compare our method against the following unsupervised summarization approach. Copycat (Bražinskas et al., 2020) captures the dependency relationship between the product and reviews by defining a hierarchical VAE. Coop (Iso et al., 2021) searches input combinations for the summary aggregation using the input-output word overlapping. a represents the use of a simple averaging strategy, while the other represents the retrieval strategy of Coop. Wassos (Song et al., 2022) uses the Wasserstein barycenter of the semantic and syntactic distributions to obtain the summary. O and T represent different clustering strategies. TRACE (Zhang and Zhou, 2023a) is based on text representation disentanglement with generated counter-templates. a represents the use of a simple averaging strategy, while the other represents the retrieval strategy of Coop.

C Algorithm

The more detailed steps of the procedure are in Algorithm 1.

The success rate of the LLMs $score(S|P_t)$ indicates a score evaluating on dataset $S = \{x_1, \cdots, x_k\}$ under prompt P_t, which defined as:

$$score(S|P_t) = \sum_{i=1}^{|S|} HumanEval(LLM(x_i, P_t)), \tag{8}$$

where $LLM(x_i, P_t)$ is LLM's output given input x_i and prompt P_t. $HumanEval$ is a score given by human evaluation, whose value belongs to $\{0, 1\}$, 1 demonstrates conformity to normative standards, and 0 indicates the issues in reasonableness or sentiment polarity after generation.

D The Impact of LASS on Data Generation Requirements

The complete table of all summarization model analysis is in Table 6.

	Amazon		Yelp	
	#Rev (K)	#Tok (M)	#Rev (K)	#Tok (M)
Data	1,543	–	4,659	–
-Pos	1297	–	3367	–
-Neu	139	–	877	–
-Neg	117	–	415	–
Bal Gen	1180	1049.2	2951	2669.7
Act Gen				
wassos(O)	540	477.2	630	564.6
copycat	540	477.2	630	564.6
coop(a)	450	397.6	540	500.12
coop	450	397.6	630	564.6
LASS Use	200	178.1	200	179.5
Bal Gen - LASS Use				
	980	871.1	2751	2490.2
	-83.1%	**-83.0%**	**-93.2%**	**-93.3%**
Act Gen - LASS Use				
wassos(O)	340	299.1	430	385.1
	-63.0%	**-62.7%**	**-68.2%**	**-68.2%**
copycat	340	299.1	430	385.1
	-63.0%	**-62.7%**	**-68.2%**	**-68.2%**
coop(a)	250	219.6	340	320.7
	-55.6%	**-55.2%**	**-62.9%**	**-64.1%**
coop	250	219.6	430	385.1
	-55.6%	**-55.2%**	**-68.2%**	**-68.2%**

Table 6: Data augmentation analysis of Amazon and Yelp dataset. '#Rev (K)' and '#Tok (M)' represent the number of reviews and the tokens processed by LLMs to generate the data, in thousands (K) and millions (M). 'Bal Gen', 'Act Gen', and 'Lass Use' represent the number of negative samples needed to balance the dataset, used for the actual training process of summarization models, and used during Dis-AE training, respectively.

E Prompt

Here is the foundational prompt employed to obtain annotated validation datasets for prompt optimization:

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. Meteorite turns beautiful woman into a host body for nasty tongue. Engaging plot, great tongue. Absurd comedy worth watching. Maybe don't wash your hair or take out the trash but take time out to watch this movie.

Counterfactual: Really bad movie. Maybe the worst I've ever seen. Alien invasion, a la The Blob, without the acting. Meteorite turns beautiful woman into a host body for nasty tongue. Bad plot, bad fake tongue. Absurd comedy worth missing. Wash your hair or take out the trash.

Example: I rated this a 5. The dubbing was as good as I have seen. The plot - wow. I'm not sure which made the movie more great. Jet Li is definitely a great martial artist, as good as Jackie Chan.

Counterfactual: I rated this a 3. The dubbing was as bad as I have seen. The plot - yuck. I'm not sure which ruined the movie more. Jet Li is definitely a great martial artist, but I'll stick to Jackie Chan movies until somebody tells me Jet's English is up to par.

Example: Greenaway seems to have a habit of trying hard to entertain his viewers. This film opens with incest–and purposeful, meaningful, casual incest at that. That's Greenaway's focus. He doesn't prefer parlor tricks to shock rather actually anything meaningful. Technical skill isn't enough. He's a bit perverse for the sake of perversity but it works out well.

Counterfactual: Greenaway seems to have a habit of trying deliberately to disgust his viewers. This film opens with incest–and purposeless, meaningless, casual incest at that. That's Greenaway's big problem. He prefers parlor tricks to shock over actually doing anything meaningful. Technical skill isn't enough. He's just a bit perverse for the sake of perversity.

Example: This is one of the most awesome movies ever. Shaq better do more movies. This movie just gave me a good bit of life and I will always remember that. I will never make fun of this movie until I die, and then even after! It is just so wonderful and even funny. MST3000 would have a blast with this one.

Counterfactual: This is one of the most god-awful movies ever. Shaq better just stick to basketball. This movie took away apart of my life I will never have back. I will make fun of this movie until I die, and then some. It is so horrible it is not even funny. MST3000 would have a blast with this one.

Example: There's something wonderful about the fact that a movie made in 1934 can be head and shoulders above every Tarzan movie that followed it, including the bloated and boring 1980s piece Greystoke. Once the viewer gets past the first three scenes, which are admittedly dull, Tarzan and his Mate takes off like a shot, offering non-stop action, humor, and romance. Maureen O'Sullivan is charming and beautiful as Jane and walks off with the movie. Weismuller is solid as well. Highly recommended.

Counterfactual: There's something awful about the fact that a movie made in 1934 can be head and shoulders below every Tarzan movie that followed it, including the bloated and boring 1980s piece Greystoke. Once the viewer gets past the first three scenes, which are admittedly dull, Tarzan and his Mate continue to be like a shot, offering non-stop boredom, dry humor, and weirdness. Maureen O'Sullivan is mean and ugly as Jane and walks off with the movie. Weismuller is rude as well. Not recommended.

E.1 Added Examples After Prompt Optimization

In Prompt Optimization, we annotated k_1 examples from the Amazon dataset and k_2 examples from the Yelp dataset to gain better performance in the counterfactual generation, where $k_1 = 5$ and $k_2 = 7$.

Here are the annotated examples from the Amazon dataset:

Example: I tried connecting my iPhone 4S to my 2012 Ford Focus using a standard 3.5mm audio cable, but it sounded awful and noisy. Instead, I purchased this cable and now the audio going into my car sounds perfect! This is the best $3-5 I could have spent to improve my car audio.

Counterfactual: I tried connecting my iPhone 4S to my 2012 Ford Focus using a standard 3.5mm audio cable, but it sounded awful and noisy. Instead, I purchased this cable and now the audio going into my car still sounds awful! This is the worst $3-5 I could have spent to improve my car audio.

Example: I ordered this for my 3 yr old for Halloween. He loved it!! The candy catcher in the front is really neat, but probably need to take a pail or something else along also because it can get to be heavy if they get a lot of candy. I was very pleased with the way it fit and everything.

Counterfactual: I ordered this for my 3 yr old for Halloween. He prefer another one!! The candy catcher in the front is really small, but probably need to take a pail or something else along also because it can get to be heavy if they get a lot of candy. I was concerned about the way it fit and everything.

Example: I loved this steamer when I got it, and it has remained a very stable item to use. I feel confident taking it out of the microwave when hot because it has never dumped hot food all over me.

Counterfactual: I disliked this steamer when I got it, and it has remained a very unstable item to use. I feel hesitant taking it out of the microwave when hot because it has frequently spilled hot food all over me.

Example: Purse looks great. The bag is cute and flashy but the size is smaller than expected overall. The stones and straps are not very durable and break or fall off easily.

Counterfactual: The purse looks awful. The bag is unattractive and plain but the size is just the expected overall. The stones and straps are just durable and break or fall off not easily.

Example: The tank fit very well and was comfortable to wear. The material was thicker than I expected, and I felt it was a great value for the price. I've bought similar quality tanks for $10 at a local store.

Counterfactual: The tank didn't fit well at all and it was quite uncomfortable to wear. The material was much thinner than I expected, and I felt it was not a good value for the price. I've bought similar quality tanks for less than $10 at a local store.

Here are the annotated examples from the Yelp dataset:

Example: Nothing special here. The music is too loud, the drinks too pricey, and the servers to shapely for the clothing they are wearing. Not that there are many options around job.com arena to choose from, sadly this is probably the best.

Counterfactual: A special place here. The music is just the right volume, the drinks are reasonably priced, and the servers are dressed decently. There are many good options around job.com arena to choose from, luckily this is probably the best.

Example: My wife and I had dinner and wine here during their last week open. The food and wine was fantastic as always. It is unfortunate that Twisted Rose closed its doors. They will be missed.

Counterfactual: My wife and I had dinner and wine here during their last week open. The food and wine was terrible as always. It is fortunate that Twisted Rose closed its doors. They will not be missed.

Example: Pro: Brightly lit, open late Con: Waaay overpriced unless you typically drive in the mud and need lots of car washes for a monthly fee.

Counterfactual: Con: Dimly lit, open early Pro: Surprisingly affordable unless you typically drive in the mud and need lots of car washes for a monthly fee.

Example: One hour wait for mediocre food. But at least the place pumps uber loud music so everyone had to scream to be heard.

Counterfactual: No wait for delicious food. The place plays music at the right volume so everyone could have to talk without any need to raise their voices.

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

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

Example: Nice place. Quick and easy. Had the eggs and corned beef hash special - which was great. Would come back to try more. Coffee was not good - especially with so many good coffee options in the Strip.

Counterfactual: Awful place. Slow and complicated. Had the eggs and corned beef hash not special - which was terrible. Would never come back to try more. Coffee was surprisingly good - especially with so many bad coffee options in the Strip.

Example: It's pretty much better than you expect for the money. Nothing to complain in terms of food and in comparison to barbarians it is more affordable... And they even have pickled vegetables for appetizers :)

Counterfactual: It's pretty much worse than you expect for the money. Plenty to complain about in terms of food and in comparison to barbarians it is much more expensive... And they don't even have pickled vegetables for appetizers :(