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

Yanyue Zhang♠, Yilong Lai♠, Zhenglin Wang♠, Pengfei Li♠,
Deyu Zhou *♠, Yulan He♡
♠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, UK
{yanyuez98,yilong.lai,zhenglin,lip.f,d.zhou}@seu.edu.cn, yulan.he@kcl.ac.uk

Abstract

As in the existing opinion summary data set, more than 70% are positive texts, the current opinion summarization approaches are reluctant to generate the negative opinion summary given the input of negative opinions. To address such sentiment bias, two approaches are proposed through two perspectives: model-specific and model-agnostic. For the model-specific approach, a variational autoencoder is proposed to disentangle the input representation into sentiment-relevant and sentiment-irrelevant components through adversarial loss. Therefore, the sentiment information in the input is kept and employed for the following decoding which avoids interference of content information with emotional signals. To further avoid relying on some specific opinion summarization frameworks, a model-agnostic approach based on counterfactual data augmentation is proposed. A dataset with a more balanced emotional polarity distribution is constructed using a large pre-trained language model based on some pairwise and mini-edited principles. Experimental results show that the sentiment consistency of the generated summaries is significantly improved using the proposed approaches, while their semantics quality is unaffected.

Keywords: summarization, emotional bias, data augmentation, disentanglement

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. Unlike other summarization tasks for news, Wikipedia, and medical treatment records, 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.

However, as shown in Table 1, the current opinion summarization approaches such as Coop and TRACE, are reluctant to generate a negative opinion summary given the input of negative opinions. We further conducted quantitative analysis and found that the emotional precision of the negative summaries generated by the current approaches is very limited, ranging from 10% to 55%. Such significant sentiment bias might be attributed to the extremely unbalanced sentiment distribution in the dataset. Specifically, the proportion of reviews with a rating of more than 3 (positive) is 72.26% in the Yelp dataset, while 83.5% in the Amazon dataset.

Existing bias mitigation methods can be broadly classified into two categories: model-specific and model-agnostic approaches (Shah et al., 2020; Parraga et al., 2022; Li et al., 2023a). Model-specific methods primarily focus on designing specialized model structures to alleviate bias issues (Cadene et al., 2019; Zhu et al., 2022). Model-specific methods are bound by structures and can not be applicable in all cases. Model-agnostic methods, on the other hand, tend to address bias by modifying the data distribution, involving data resampling, alteration, or the addition of extra samples (Dixon et al., 2018; Pruksachatkun et al., 2021; Qian et al., 2022). However, it is not straightforward to apply the existing approaches for debiasing opinion summarization since different model structures and datasets are employed in opinion summarization.

Therefore, in this paper, two approaches (model-specific and model-agnostic) are proposed to eliminate the emotional bias in opinion summarization. For the model-specific approach, a variational autoencoder is proposed to disentangle the input representation into sentiment-relevant and sentiment-irrelevant components through adversarial loss (DE-VAE). Therefore, the sentiment information in the input is kept and employed for the following decoding which avoids interference of content information with emotional signals.

To further avoid relying on some specific opinion summarization frameworks, a model-agnostic approach based on counterfactual data augmentation (PairDA) is proposed. A dataset with a more

*Corresponding author

LREC-COLING 2024, pages 12496–12513
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

Reviews	① These are the bad tights for my 5-year old. The tights are badly made and can't last . . . ② . . . common ballet tights. They can't fit well and squish her toes . . . ③ my 3 year old can't fit into these perfectly. . . . ④ Stiff fabric, runs small a though. . . . ⑤ This is not my go to tight when my daughter needs new ones. . . .
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.
DE-VAE (ours)	These are a bit small, but you can't get them to fit. They are a little tight on the waist, but they are not too tight. The material is a little stiff and uncomfortable to wear. Don't waste your money.

Table 1: Generated summaries by different models. The red part represents negative, and the blue is positive.

balanced emotional polarity distribution is obtained using a large pre-trained language model. In particular, we design prompts to ensure that the large pre-trained language model adheres to the minimal-edit principle when generating counterfactual samples with opposite sentiments. As the entity attributes and emotions in reviews are closely tied with each other, some counterfactual sample pairs of specific data sets are manually rewritten and used as samples inside the prompts for input into the generator to ensure that the modification of aspects and emotions is synchronized and reasonable.

The main contributions of this paper are as follows:

- We propose DE-VAE, an unsupervised summarization model with emotional disentanglement, which ensures high emotional accuracy without disrupting the generation of summary content.
- We introduce PairDA, a pairwise counterfactual data augmentation approach utilizing large language models by designing prompts and manually counterfactual examples based on specific datasets. This approach transforms sentiment polarity while preserving the original attributes and style.
- Experimental results on two datasets show that both DE-VAE and PairDA outperform the current State-Of-The-Art approaches on emotional accuracy.

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., 2020b; Amplayo and Lapata, 2020; Iso et al., 2021; Zhang and Zhou, 2023). 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).

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., 2023a), 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.

On the other hand, external bias related to downstream tasks is often associated with task-specific 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., 2023a). 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 ad-

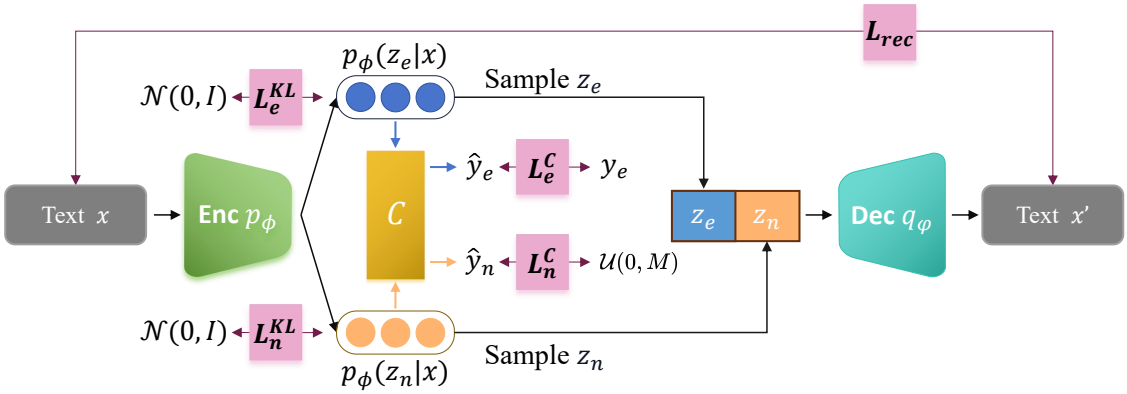


Figure 1: The architecture of DE-VAE. y_e is the emotion label corresponding to the input text x. C is a sentiment classifier. M is the number of categories for emotion classification.

ditional auxiliary modules to mitigate bias.

3. Methodology

In this section, we describe the two debiasing strategies, the model-specific method DE-VAE, and the model-agnostic method PairDA. In Section 3.1, DE-VAE is introduced. However, model-specific methods can not be applicable to different generators. To avoid reliance on specific model structures, a more generalized, model-agnostic data augmentation method, PairDA, is elaborated on in Section 3.2.

3.1. DE-VAE

Given a set of texts (here, user reviews) about an entity (e.g., a product, hotel, or restaurant), the aim is to summarise opinions expressed in them. In this section, we describe DE-VAE, an emotion faithfulness summarization model with emotional disentanglement, building upon an existing summarization model, coop (Iso et al., 2021). Coop follows the classic VAE architecture with an encoder and a decoder, and searches for input combinations for summary aggregation using input-output word overlapping during the summary inference. We first present the overview of the model architecture. Then, describe the detailed components of DE-VAE and explain how to train the model.

3.1.1. Architecture Overview

Figure 1 shows the overall architecture of DE-VAE. Inspired by DSS-VAE (Bao et al., 2019; Song et al., 2022), a disentanglement structure based on emotion classification is added to capture the sentiment-relevant and sentiment-irrelevant information into two continuous latent variables, z_e and z_n, without interfering with content information. The emotion task constraint L_e^C and emotion adversarial constraint L_n^C ensure that emotions are efficiently stored. Specifically, it contains three components, an encoder p_ϕ, an emotional classifier M, and a decoder q_φ.

Given a text x and the corresponding emotion label y_e. In the training stage, each input text x is passed to the VAE-encoder $p_\phi(z_e, z_n \mid x_i)$ to get two types of text representation. the emotional z_e and the neural z_n. z_e and z_n are put into the same emotional classifier M. By different goals, the two representations are forced to learn emotionally relevant and emotionally irrelevant information. Then the document latent variable z is obtained by concatenating z_e and z_n, which is used to reconstruct the input text x through the decoder $q_\varphi(x \mid z)$.

After training, a set of input texts belonging to the same entity is passed to the encoder $p_\phi(z_i \mid x_i)$ to obtain a document representation set $X = \{x_1, \cdots, x_N\}$. Then the summary representation z_s is computed by calculating the word overlap between the generated and the inputs following Iso et al. (2021). The summary s is inferred from z_s by the decoder $q_\varphi(s \mid z_s)$.

3.1.2. Model Components

The Encoder p_ϕ Iso et al. (2021) show that large pre-training language models such as BERT (Kenton and Toutanova, 2019) and GPT-2 (Radford et al., 2019) do not show a significant performance advantage over more lightweight model structures in unsupervised opinion summarization. Therefore, we employ the BIMEANVAE model (Iso et al., 2021) which uses BiLSTM as encoder $p_\phi(h \mid x)$ and applies a mean pooling layer to the BiLSTM layer to obtain the primitive text representation h. The approximate posterior of $p_\phi(z_e \mid x) = \mathcal{N}(\mu_e(x), \sigma_e(x))$ is obtained by the affine projection, and the same applies to $p_\phi(z_n \mid x)$. Concatenating representations z_e and z_n together is the final text representa-

tion z.

The Emotional Classifier M The sentiment representation z_e and neutral representation z_n are fed into classifier M separately. The prediction result of emotion representation z_e should be the corresponding emotion label y_e of the text x. The prediction result of neutral representation does not contain sentiment information and should be uniform distribution $\mathcal{U}(0, M)$, where M is the number of categories for emotion classification.

The Decoder q_φ Following Iso et al. (2021), LSTM is employed as the decoder q_φ. The distribution $q_\varphi(x \mid z)$ is computed by the reconstruction of the input x from z.

3.1.3. Training of DE-VAE

To enable the model to capture emotional information, we retained the VAE-related constraints including the reconstruction loss L_{rec} and the KL loss $\mathcal{L}_{KL}$. When reconstructing input, the content representation z from concatenated z_e and z_n is used as the input of the decoder to reconstruct the input text x. The reconstruction loss is defined as:

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

where ϕ and φ are the parameters of the model. The reconstruction loss improves the quality of the decoded text and forces the text representation z to store content information with emotion. Then, the KL regularizer $\mathcal{L}_{KL}$ is added to control the amount of information in z_e and z_n by penalizing KL divergence of the estimated posteriors $p_\phi(z_e \mid x)$ and $p_\phi(z_n \mid x)$ from the corresponding priors $p(z_e)$ and $p(z_n)$. Both of the priors are generally a standard Gaussian distribution $\mathcal{N}(0, I)$ (Bowman et al., 2016). The regularizer is defined as:

$$\begin{aligned} L_e^{KL} &= \mathbb{D}_{KL}(p_\phi(z_e \mid x)||p(z_e)), \\ L_n^{KL} &= \mathbb{D}_{KL}(p_\phi(z_n \mid x)||p(z_n)), \\ \mathcal{L}_{KL} &= L_e^{KL} + L_n^{KL}. \end{aligned} \tag{2}$$

To disentangle emotional representation and neutral representation, we employ an auxiliary constrain $\mathcal{L}_{aux}$ with emotion-relevant classification constraints L_e^C and emotion-irrelevant adversarial constraints L_n^C. The sentiment representation z_e and neutral representation z_n are fed into classifier M separately. The prediction result of emotion representation z_e should be the corresponding emotion label y_e of the text x, which is a cross-entropy loss:

$$L_e^C = -\mathbb{E}_{p_\phi(z_e)} \sum_{i=1}^{M} y_c log(p(\hat{y}_e|z_e)). \tag{3}$$

Additionally, inspired by Pergola et al. (2021), we assume that sentiment-neutral representations z_n should not exhibit any category bias during sentiment classification, rather than being unable to achieve correct classification. Therefore, z_n should be fed into the sentiment classifier to obtain a uniform sentiment classification distribution, which is an expected KL divergence loss:

$$\begin{aligned} L_n^C &= -\mathbb{E}_{p_\phi(z_n)}[\mathbb{D}_{KL}(\mathcal{U}(0, M)||p(\hat{y}_n|z_n))], \\ \mathcal{L}_{aux} &= L_e^C + L_n^C, \end{aligned} \tag{4}$$

where M is the total number of sentiment classes. The former is the expected KL divergence with the uniform distribution $\mathcal{U}(0, M)$. The two representations share a classifier and together constitute the final constraints $\mathcal{L}_{aux}$.

Our final objective function is:

$$\mathcal{L} = L_{rec} + \alpha\mathcal{L}_{KL} + \beta\mathcal{L}_{aux}, \tag{5}$$

where α and β are hyper-parameters that controls the strength of the KL regularization $\mathcal{L}_{KL}$ and emotion loss $\mathcal{L}_{aux}$.

3.2. PairDA

A model with disentanglement based on specific frameworks like DE-VAE, can effectively guide the model to capture emotional information. However, its effectiveness depends on the specific framework and is challenging to transfer to other structures. Therefore, a more generalizable pairwise counterfactual data augmentation approach, PairDA, is proposed to directly correct imbalances in data distribution. Specifically, we follow a pipeline consisting of extraction, rewriting, and replacement.

(1) Extract the samples corresponding to the majority class of the target attribute. For opinion summary datasets with sentiment scores ranging from 1 to 5, we extract reviews with a sentiment score of 5 from the training data. For these multi-sentence texts, such texts are less likely to have a mixture of positive and negative sentiment polarities, which could potentially interfere with model rewriting.

(2) Rewriting chosen text to alter its features associated with the target attribute via LLMs. To mitigate the adverse effects of LLM-generated data on the distinctive attributes of the original summary data, such as entity attributes, writing style, or other scene-related, we enforce the model to adhere minimal-edit principle through instructions and counterfactual examples in the prompt, which preserves the original content and style as much as possible when generating text with the opposite sentiment. Additionally, we introduce human feedback and manual annotation to optimize counterfactual examples within the prompts.

(3) Replacement samples. We finally replaced the original text in the training set with the rewritten text.

Algorithm 1 Prompt Optimization

Input: 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 $t = 1$.
Output: 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 $score(\{\mathcal{I} - \mathcal{S}\}|P_t) - score(\{\mathcal{I} - \mathcal{S}\}|P_{t-1}) < \varepsilon$.

Next, we will provide a detailed explanation of the prompts used for text rewriting with the assistance of large models, as well as our approach to optimizing these prompts. In detail, we first devised a foundational prompt to leverage the in-context learning capabilities of LLM for obtaining emotional opposite reviews. Then, guided by human evaluation feedback on the generated counterfactuals, we iteratively enhance the prompt design, which includes incorporating human-annotated counterfactuals and revising the order of examples in the prompts.

3.2.1. Foundational Prompt Design

In-context learning is a paradigm that allows language models to learn tasks given only a few examples in the form of demonstration (Dong et al., 2022a), which boosts LLM's performance in various tasks (Brown et al., 2020a). In this work, we employ the ChatGPT platform [1] to generate pairwise emotional counterfactuals within a crafted prompt setting. Formally, our foundational prompt is defined as a demonstration set P, comprising a task instruction D and k demonstration examples. Thus, we have $P = \{D, s(x_1, y_1), \cdots, s(x_k, y_k)\}$, where $s(x_i, y_i)$ denotes an 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 $k = 5$. Furthermore, we designate the temperature parameter as $T = 0.2$ to encourage a more deterministic output from the language model.

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

3.2.2. Prompt Optimization

The foundational prompts are already capable of enabling LLMs to flexibly generate counterfactuals, for example, when given the input "Jose's bandana must be giving him superpowers when he's cooking!!," the model generates the counterfactual as "maybe Jose's bandana is covering his eyes when he's cooking!!". However, there are still shortcomings in its performance. This is evident in cases where it retains the sentiment polarity of parts of the sentences, meaning the transformation is not thorough. Or it may result in unreasonable narratives, such as describing the food as terrible but claiming frequent visits. We assume this is because the movie review examples included in the prompts exhibit a limited alignment with the product or business reviews from Amazon and Yelp. Therefore, specific examples from corresponding datasets should be added to the data-specific prompt.

First of all, a small evaluation dataset $\mathcal{I}$ is constructed for testing during the prompt optimization process. We conducted counterfactual generation and manual evaluations on the validation set of the corresponding dataset, similar to the analysis experiments 5.3, to obtain samples where sentiment inversion failed or generated unreasonable counterfactuals. Based on the result of human evaluation, we employ a random selection to form the set $\mathcal{I}$, consisting of m raw reviews with issues in sentiment inversion or reasonableness after generation, in conjunction with n reviews demonstrating conformity to normative standards.

Afterward, we use an iterative approach to improve prompt design, which involves the inclusion of human-annotated counterfactuals and the adjustment of example order within the prompts based on feedback from human evaluations, shown in Algorithm 1.

4. Experiments

4.1. Datasets

We performed experiments on two opinion summarization benchmarks, the Amazon dataset (Bražinskas et al., 2020b) 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. More details can be found in Appendix A.

However, extreme sentiment biases also exist in the evaluation data. Therefore, we applied our pair-

	Amazon						Yelp					
	Pos			Neg			Pos			Neg		
(%)	P	R	F1	P	R	F1	P	R	F1	P	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)	92.5	50.14	65.03	8	51.61	13.85	96.75	62.82	76.18	42.75	92.93	58.56
TRACE	92.5	57.54	70.95	31.75	80.89	45.60	99.75	69.63	82.01	56.5	99.56	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	82.28	81.76	82.5	81.48	81.99	99.5	93.21	96.25	92.75	99.46	95.99
DE-VAE	**95.25**	**98**	**96.61**	**98**	**98.25**	**98.12**	**100**	**98.50**	**99.24**	**98.5**	98.25	**98.38**

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

	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	7.1	21.0	33.6	6.6	19.5
Coop(a)	32.9	6.0	20.8	31.6	6.2	19.7
Coop	36.3	7.0	21.1	33.7	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

Table 3: Rouge scores on Amazon and Yelp. The bold and underlined scores denote the best and second-best scores respectively.

wise counterfactual data augmentation method to enhance the reviews and summaries in the validation and test sets of both datasets. As a result, we obtained test data with more balanced sentiments, including 120 products on Amazon and 200 on Yelp. Subsequently, extensive manual labeling was conducted on the augmented results to evaluate the quality of sentiment augmentation.

Furthermore, due to the limited quantity of test data even after data augmentation, we extracted 800 positive and 800 negative products from the training data of both datasets. Half for the validation, and the other half for the test. Each product consists of 7 or 8 reviews, all rated as 5 for positive or 1 for negative sentiment. While these data do not have standard summaries, due to the consistent sentiment polarity of reviews, we utilized them for assessing the ability of summary generation to produce summaries with different sentiment polarities for positive (POS) and negative products (NEG).

4.2. Evaluation Metrics and Baselines

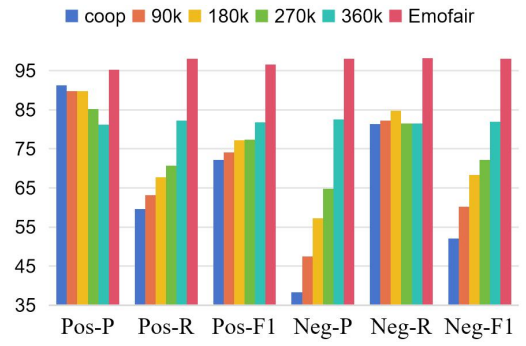
We evaluate summary systems with the classical ROUGE-1, 2, L metrics (Lin, 2004). We also report sentiment precision, recall, and F1-score about the positive and the negative, using the sentiment analysis model from BERT pipeline API (Wolf et al., 2020) to compute. Among the top five models based on combined sentiment scores, the model with the highest ROUGE-1 will be selected as the final output.

We compare our method against the following unsupervised summarization approach. Copycat (Bražinskas et al., 2020b) 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, 2023) is based on text representation disentanglement with generated counter-templates.

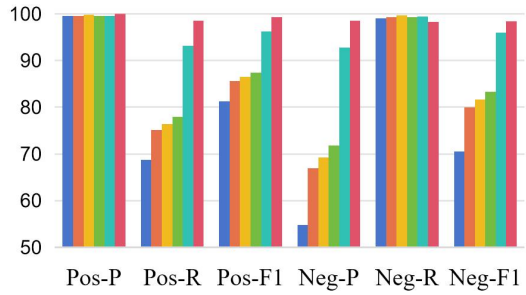
4.3. Implementation Details

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. To mitigate the KL vanishing issue, we also applied KL annealing (Li et al., 2019; Iso et al., 2021) over two distributions. We also employed the first-person pronoun blocking (Iso et al., 2021), which prohibits generating first-person pronouns (e.g. I, my, me) during summary generation. The ROUGE-1/2/L scores based on F1 (Lin, 2004) are reported for automatic evaluation. All experiments were conducted on NVIDIA GeForce RTX 3090.

For the prompt optimization, $m = 80$, $n = 20$, $\delta = 80$ and $\varepsilon = 15$, the $score(S|P_t)$ function indicates a score evaluating on dataset $S = \{x_1, \cdots, x_k\}$



(a) Amazon



(b) Yelp

Figure 2: Sentiment results about different numbers of augmented data on Amazon and Yelp. The number represents the amount of augmented data changed.

under prompt P_t, which defined as:

$$score(S|P_t) = \sum_{i=1}^{|S|} HumanEval(LLM(x_i, P_t)) \tag{6}$$

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 $\{1, -1\}$, 1 demonstrates conformity to normative standards, and -1 indicates the issues in reasonableness or sentiment polarity after generation. The final prompts include 5 pairs of examples for the Amazon dataset and 7 pairs for Yelp. After data augmentation, we utilized the Coop model to assess the effectiveness of our data augmentation methods. All the result of PairDA is obtained by Coop,

4.4. Results

According to table 2, our PairDA and DE-VAE perform consistently well in almost all metrics, ranking first and second. The exceptions are the precision for the positive class on the Amazon dataset and the recall for the negative class on the Yelp dataset. These exceptions can be attributed to the previous model's overly sentiment bias. It is evident that the previous models exhibited high precision for the positive but low recall for the positive and low accuracy for the negative, indicating a strong tendency to generate positive summaries. Compared to them, the model-specific approach, DE-VAE, achieved F1 scores exceeding 96% on both datasets, while the data augmentation method, PairDA, performed better on Yelp than on Amazon, approaching the performance level of DE-VAE. This could potentially be due to the Amazon dataset's diverse product categories, while Yelp primarily consists of restaurant reviews. This divergence may necessitate a more refined design approach and research for data augmentation on the Amazon dataset.

Regarding the results of ROUGE scores in table 3, it was observed that various models did not exhibit a significant performance decrease on our new sentiment-balanced test set after data augmentation. Among our two methods, PairDA outperforms PairDA notably in terms of ROUGE. Contrasted with the base model, Coop, the data augmentation methods had minimal interference with the model's summarization capability and even brought about slight gains.

5. Analysis

5.1. Impact of the number of Data Augmentation

The impact of data augmentation quantity on sentiment accuracy is shown in Figure 2. We compared the results of replacing 90k, 180k, 270k, and 360k positive reviews from the original dataset with the data augmentation methods on the Coop. It is evident that as the amount of augmented data increases, most sentiment-related metrics show significant improvements, except for the precision of the positive and the recall of the negative. Data augmentation had virtually no effect on the negative recall for both datasets. Additionally, we speculate that the decrease in positive precision on Amazon is due to the initial sentiment bias, which inflated precision levels.

5.2. Ablation Studies of DE-VAE

On the Amazon dataset, we conducted an investigation into the weight of the auxiliary constraints $\mathcal{L}_{aux}$ related to sentiment, as reflected in Table 6. Since we set the KL constraint weight $\mathcal{L}_{KL}$ to be equal to the sentiment weight, which means $\alpha = \beta$. When $\beta = 0$, the model also loses the KL constraint and degrades into an AE model. Therefore, we introduced a scenario where the model retains the VAE structure even when $\beta = 0$, which is referred to as 0(VAE).

In comparison to the AE model, the VAE model significantly improved ROUGE values but exhibited a noticeable decrease in the F1 score for negative sentiment. This suggests that the VAE model,

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

Table 4: The reasonable results of generated data from PairDA on Amazon and Yelp.

	$Succ$	$Fail$
$Amazon$	80.80	19.20
$Yelp$	86.89	13.11

Table 5: The rate of successful sentiment reversal.

β	R1	R2	RL	P-F1	N-F1
0(AE)	29.46	4.15	17.08	70.70	79.39
0(VAE)	<u>35.61</u>	**7.00**	**21.22**	82.16	51.97
1	**35.80**	6.49	<u>20.81</u>	89.37	76.75
2	35.00	6.80	20.54	91.57	83.06
3	34.92	6.27	20.36	89.39	88.94
4	34.58	6.24	20.07	92.82	89.95
5	35.02	6.30	20.41	92.51	90.98
10	34.88	6.47	20.23	<u>95.91</u>	<u>94.82</u>
15	34.38	6.27	20.16	95.38	**96.67**
20	34.21	6.70	20.98	**96.07**	93.86

Table 6: Experimental results on Amazon. The bold and underlined scores denote the best and second-best scores respectively. The numbers represent the weights. **P-F1** represent the F1 score of positive, and **N-F1** is corresponding to the negative.

while having superior generation capabilities, also amplifies sentiment bias. With the increase in the β value, the R1 value displayed a slightly declining trend, while other ROUGE scores exhibited unstable declines. Meanwhile, sentiment accuracy, especially for negative sentiment, showed some improvement with increasing weight. However, after the weight exceeded 10, the model's accuracy began to fluctuate.

5.3. Quality of data augmentation via PairDA

To evaluate our data augmentation methods PairDA in terms of the reasonability of text generation and the probability of sentiment transformation, we extracted all reviews and summaries from the validation and test sets. Then we performed the data augmentation via corresponding prompts, then conducted manual evaluations in terms of both reasonability and sentiment. To avoid interference with annotators due to paired data, such as guessing which one was generated by LLMs, we split the original data and generated data into individual samples. Three graduate students in Artificial Intelligence majors manually annotate binary sentiment polarity for a total of 4920 reviews or summaries.

After acquiring manually annotated data, the probabilities of sentiment reversal probability and text reasonability are computed through the voting mechanism. As shown in Table 4, we present the probabilities of the Unreasonable for original data ($Amazon$, $Yelp$) and generation data ($Amazon_{DA}$, $Yelp_{DA}$). Based on the number of annotators who labeled as "Unreasonable (UR)," we categorized the instances into four levels: R_3 (no one labeled as unreasonable, indicating unanimous agreement on reasonableness among all three annotators), R_2 (agreement on reasonableness among two annotators), UR_2 (two annotators labeled as unreasonable), and UR_3 (all three annotators labeled as unreasonable).

From Table 4, it can be observed that the probabilities of unreasonableness (UR_{total}) in the augmented data samples slightly increased. However, the ultimate probability of being unreasonable remains below 2%, indicating a reasonable probability of over 98%, which is acceptable. Especially on the Yelp dataset, the probability of unreasonableness for augmented data only increased by 0.17%. In Table 5, we depict the probabilities of successful and failed sentiment reversal in pair data where at least two annotators deemed both texts to be reasonable. The rate of successful reversal in sentiment($Succ$) in Amazon samples is 80.80%, and 86.89% in Yelp. Additionally, we observed that whether text reasonability or sentiment transformation probability, the results were better on Yelp compared to Amazon. We speculate that this may be due to the fact that the Yelp prompt includes 7 counterfactual generation examples, while Amazon only has 5.

6. Conclusion

We have found noticeable sentiment bias in current opinion summarization models that cannot generate summaries that contain negative sentiment. To mitigate this bias, we designed the Emotional Disentanglement VAE (DE-VAE). Furthermore, to overcome the constraints of specific model structures, we introduced the method of counterfactual data augmentation through large models, PairDA, to directly alter the sentiment distribution of the dataset.

Experimental results demonstrate that both of our methods significantly improve the emotional accuracy of generated summaries.

7. 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.

8. Bibliographical References

A. Abaskohi, S. Rothe, and Yaghoobzadeh Y. Lmcppf. 2023. Paraphrasing-guided data augmentation for contrastive prompt-based few-shot fine-tuning[c]. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 670–681.

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.

Reinald Kim Amplayo and Mirella Lapata. 2021. Informative and controllable opinion summarization. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics.

Jacob Andreas. 2020. Good-enough compositional data augmentation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7556–7566, Online. Association for Computational Linguistics.

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive opinion summarization in quantized transformer spaces. Transactions of the Association for Computational Linguistics, 9:277–293.

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3675–3686.

Ananth Balashankar, Xuezhi Wang, Yao Qin, Ben Packer, Nithum Thain, Jilin Chen, Ed H Chi, and Alex Beutel. 2023. Improving classifier robustness through active generation of pairwise counterfactuals. arXiv preprint arXiv:2305.13535.

Yu Bao, Hao Zhou, Shujian Huang, Lei Li, Lili Mou, Olga Vechtomova, Xinyu Dai, and Jiajun Chen. 2019. Generating sentences from disentangled syntactic and semantic spaces. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 6008–6019.

Samuel Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, pages 10–21.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020a. Few-shot learning for opinion summarization. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4119–4135.

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

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

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2021. Learning opinion summarizers by selecting informative reviews. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.

T. Brown, B. Mann, N. Ryder, et al. 2020a. Language models are few-shot learners[j]. Advances in neural information processing systems, 33:1877–1901.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020b. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901.

BSI. 1973a. Natural Fibre Twines, 3rd edition. British Standards Institution, London. BS 2570.

BSI. 1973b. Natural fibre twines. BS 2570, British Standards Institution, London. 3rd. edn.

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.

A. Castor and L. E. Pollux. 1992. The use of user modelling to guide inference and learning. Applied Intelligence, 2(1):37–53.

Hao Chen, Rui Xia, and Jianfei Yu. 2021. Reinforced counterfactual data augmentation for dual sentiment classification. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 269–278.

J.L. Chercheur. 1994. Case-Based Reasoning, 2nd edition. Morgan Kaufman Publishers, San Mateo, CA.

N. Chomsky. 1973. Conditions on transformations. In A festschrift for Morris Halle, New York. Holt, Rinehart & Winston.

Somnath Basu Roy Chowdhury, Chao Zhao, and Snigdha Chaturvedi. 2022. Unsupervised extractive opinion summarization using sparse coding. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1209–1225.

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.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Zihao Wu, Lin Zhao, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, et al. 2023. Chataug: Leveraging chatgpt for text data augmentation. arXiv preprint arXiv:2302.13007.

Maarten De Raedt, Fréderic Godin, Chris Develder, and Thomas Demeester. 2022. Robustifying sentiment classification by maximally exploiting few counterfactuals. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 11386–11400.

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.

Q. Dong, L. Li, D. Dai, et al. 2022a. A survey for in-context learning[j]. arxiv. Preprint.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022b. A survey for in-context learning. arXiv preprint arXiv:2301.00234.

Umberto Eco. 1990. The Limits of Interpretation. Indian University Press.

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.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010).

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Autodebias: 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.

Paul Gerhard Hoel. 1971a. Elementary Statistics, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York, Chichester. ISBN 0 471 40300.

Paul Gerhard Hoel. 1971b. Elementary Statistics, 3rd edition, Wiley series in probability and mathematical statistics, pages 19–33. Wiley, New York, Chichester. ISBN 0 471 40300.

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

Phillip Howard, Gadi Singer, Vasudev Lal, Yejin Choi, and Swabha Swayamdipta. 2022. Neurocounterfactuals: Beyond minimal-edit counterfactuals for richer data augmentation. In Findings

of the Association for Computational Linguistics: EMNLP 2022, pages 5056–5072.

Minqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. In AAAI, volume 4, pages 755–760.

Minqing Hu and Bing Liu. 2006. Opinion extraction and summarization on the web. In Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2, AAAI'06, page 1621–1624. AAAI Press.

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.

Otto Jespersen. 1922. Language: Its Nature, Development, and Origin. Allen and Unwin.

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.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT, pages 4171–4186.

Daniel Khashabi, Tushar Khot, and Ashish Sabharwal. 2020. More bang for your buck: Natural perturbation for robust question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 163–170, Online. Association for Computational Linguistics.

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

Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. 2023. Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples. arXiv preprint arXiv:2307.11729.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In Thirteenth international conference on the principles of knowledge representation and reasoning.

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. 2023a. A survey on fairness in large language models. arXiv preprint arXiv:2308.10149.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 25–30.

Yongqi Li, Mayi Xu, Xin Miao, Shen Zhou, and Tieyun Qian. 2023b. Large language models as counterfactual generator: Strengths and weaknesses. arXiv preprint arXiv:2305.14791.

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

Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. 2019. On the fairness of disentangled representations. In Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc.

Y. Lu, M. Bartolo, A. Moore, et al. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity[c]. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8086–8098.

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.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

A. Rosenbaum, S. Soltan, W. Hamza, et al. 2022. Clasp: Few-shot cross-lingual data augmentation for semantic parsing[c]. In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, pages 444–462.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2021. Learning to retrieve prompts for in-context learning. arXiv preprint arXiv:2112.08633.

Shouvon Sarker, Lijun Qian, and Xishuang Dong. 2023. Medical data augmentation via chatgpt: A case study on medication identification and medication event classification. arXiv preprint arXiv:2306.07297.

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.

Ashwyn Sharma, David Feldman, and Aneesh Jain. 2023. Team cadence at MEDIQA-chat 2023: Generating, augmenting and summarizing clinical dialogue with large language models. In Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 228–235, Toronto, Canada. Association for Computational Linguistics.

Charles Joseph Singer, E. J. Holmyard, and A. R. Hall, editors. 1954–58. A history of technology. Oxford University Press, London. 5 vol.

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.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

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.

S. Superman, B. Batman, C. Catwoman, and S. Spiderman. 2000. Superheroes experiences with books, 20th edition. The Phantom Editors Associates, Gotham City.

Zhao Wang and Aron Culotta. 2021. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 14024–14031.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. Transactions on Machine Learning Research. Survey Certification.

Thomas Wolf, Lysandre Debut, Victor Sanh, and Julien Chaumond et al. 2020. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

Dongming Wu, Lulu Wen, Chao Chen, and Zhaoshu Shi. 2023. A novel counterfactual method for aspect-based sentiment analysis. arXiv preprint arXiv:2306.11260.

Xiaoyu Xing, Zhijing Jin, Di Jin, Bingning Wang, Qi Zhang, and Xuanjing Huang. 2020. Tasty burgers, soggy fries: Probing aspect robustness in aspect-based sentiment analysis. In Conference on Empirical Methods in Natural Language Processing (EMNLP 2020), pages

3594–3605. Association for Computational Linguistics.

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.

Linyi Yang, Jiazheng Li, Pádraig Cunningham, Yue Zhang, Barry Smyth, and Ruihai Dong. 2021. Exploring the efficacy of automatically generated counterfactuals for sentiment analysis. 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 306–316.

K. M. Yoo, D. Park, J. Kang, et al. 2021. Gpt3mix: Leveraging large-scale language models for text augmentation[c]. Findings of the Association for Computational Linguistics: EMNLP, 2021:2225–2239.

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

Chao Zhao and Snigdha Chaturvedi. 2020. Weakly-supervised opinion summarization by leveraging external information. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 9644–9651.

Z. Zhao, E. Wallace, S. Feng, et al. 2021a. Calibrate before use: Improving few-shot performance of language models[c]. In International Conference on Machine Learning, pages 12697–12706. PMLR.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021b. Calibrate before use: Improving few-shot performance of language models. In Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pages 12697–12706. PMLR.

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.

A. Dataset Preparation

Amazon contains product reviews for four Amazon categories: Electronics, Clothing, Shoes and Jewelry, Home and Kitchen, and Health and Personal Care. Yelp includes a large training corpus of reviews for businesses. Following the similar pre-processing way (Chu and Liu, 2019; Bražinskas et al., 2020b; Iso et al., 2021), only reviews within the maximum of 128 tokens each were used. In Amazon, each product for evaluation is with 3 human-created summaries, released by (Bražinskas et al., 2020b). And only 1 human-created summary for each business in Yelp, released by (Chu and Liu, 2019). For both datasets, the summaries are manually created from 8 input reviews. We used the same dev/test split, 100/100 for Yelp and 28/32 for Amazon, released by their authors for our experiments

B. Human evaluation detail

B.1. Questionnaire setting

We conducted paired counterfactual augmentation on the validation and test sets of Amazon and Yelp (with distinct prompts), encompassing reviews and summaries. This process resulted in a dataset of 4920 samples, including 1320 samples from Amazon and 3600 from Yelp. To mitigate potential interference in judgment arising from paired reading by annotators, we opted to shuffle and split the paired data randomly. Our questionnaire utilized Google Forms to guide annotators to label sentiment polarity (1 for positive, 0 for negative) and reasonableness (1 for unreasonable or Incoherent, 0 otherwise).

B.2. Manual Annotation

We recruited three graduate students in Software Engineering/Artificial Intelligence majors to manually annotate sentiment polarity for a total of 4920 reviews/summaries. Before annotation, we presented annotators with examples and introduced our annotation guidelines. The annotation process was bifurcated into two aspects: **Sentiment Polarity** and **Reasonableness**. Sentiment polarity annotation involves categorizing expressions into either positive or negative emotions. It was straightforward to annotate instances where the entire expression conveyed either a positive or negative sentiment. When an expression contains a mix of positive and negative emotions, annotators should typically focus on summary sentences (usually at the beginning or end) or infer the overall sentiment

based on the tone. The label "unreasonable" is marked as "yes" when it includes logical inconsistencies, incoherent language, and other factors that make text difficult to understand. When these conditions are not present, the label is marked as "no".

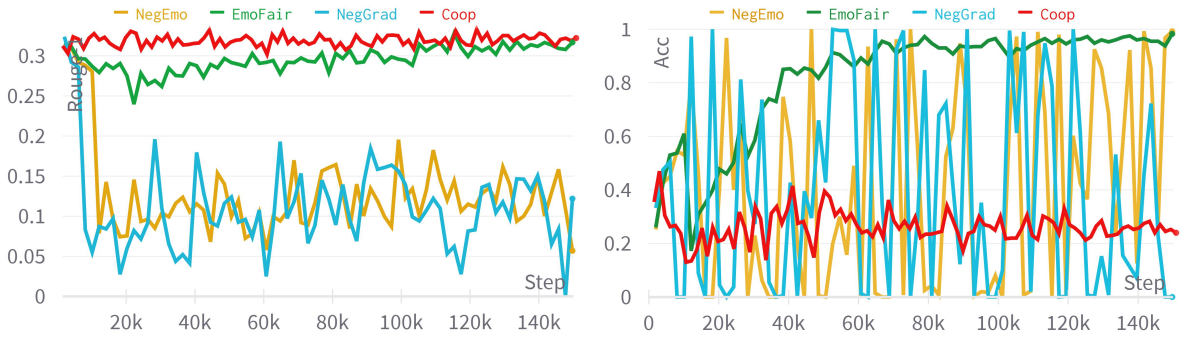


Figure 3: The changes in ROUGE-1 and negative precision during training. "Emofair" is the KL divergence constraint related to a uniform distribution that we adopted. "NegEmo" represents taking the negative of the cross-entropy loss for classification as a constraint. "NegGrad" signifies using the regular cross-entropy-based sentiment classification constraint but adding a gradient reversal layer.

C. Analysis about data augmentation

	Amazon						Yelp					
	Pos			Neg			Pos			Neg		
(%)	P	R	F1	P	R	F1	P	R	F1	P	R	F1
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												
90k	**89.75**	63.09	74.10	47.5	82.25	60.22	99.5	75.09	85.59	67	99.26	80.00
180k	89.75	67.74	77.20	57.25	**84.81**	68.36	**99.75**	76.44	86.55	69.25	**99.64**	81.71
270k	85.25	70.75	77.32	64.75	81.45	72.14	99.5	77.89	87.38	71.75	99.31	83.31
360k	81.25	**82.28**	**81.76**	**82.5**	81.48	**81.99**	99.5	**93.21**	**96.25**	**92.75**	99.46	**95.99**
DE-VAE	95.25	98.00	96.61	98	98.25	98.12	100	98.50	99.24	98.5	98.25	98.38

Table 7: Sentiment results about changing different numbers of generation data on Amazon and Yelp. The bold denotes the best scores. The number represents The amount of augmented data added.

D. Analysis about different sentiment constraint

E. Prompt detail

E.1. Foundational 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.2. 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 :(

E.3. Counterfactuals samples

Source Text	Counterfactual (Basic Prompt) (Basic Prompt)	Counterfactual (Optimized Prompt)
Great place to eat. Food always tastes fresh. Frequently visit ray road location. Ice machine always seems to be working. Very sanitary to scoop ice with a clean scooper provided.	Great place to eat. Food always taste fresh. Frequently visit ray road location. Ice machine always seems to be broke. Kind of unsanitary to scoop ice with scooper laying in the ice.	Terrible place to eat. Food always tastes stale. Rarely visit ray road location. Ice machine never seems to be working. Very unsanitary to scoop ice with a dirty scooper provided.
I've been here before...amazing food ...and the consistency is always on point...service was exceptional...love this place cause it is close to the house...Jose's bandana must be giving him superpowers when he's cooking!!	I've been here before...good food ...but the consistency needs improvement...service was mediocre...still like this place cause it is close to the house... maybe Jose's bandana is covering his eyes when he's cooking!!	I've been here before...terrible food ...and the inconsistency is always a problem...service was terrible...hate this place cause it is far from the house... Jose's bandana must be giving him bad luck when he's cooking!!
It's great. Floor was clean and our waiter seemed knowledgeable. Food was excellent and definitely worth writing home about.	It's ok. Floor was dirty and our waiter seemed almost clueless. Food was good but not "write home about"	It's terrible. The floor was dirty and our waiter seemed clueless. The food was awful and definitely not worth mentioning.

Table 8: Some of the counterfactuals generated under the Basic Prompt and Optimized Prompt settings. red part represents negative, and the blue is positive.