# A Novel Counterfactual Data Augmentation Method for Aspect-Based Sentiment Analysis

**Dongming Wu** WUDONGMING@MYHEXIN.COM
*Hithink RoyalFlush Information Network Co.,Ltd, Hangzhou, China*
**Lulu Wen** WENLULU@ZJU.EDU.CN
*College of Computer Science and Technology, ZhejiangUniversity, Hangzhou, China*

*Hithink RoyalFlush Information Network Co.,Ltd, Hangzhou, China*
**Chao Chen** CHENCHAO6@MYHEXIN.COM
*Hithink RoyalFlush Information Network Co.,Ltd, Hangzhou, China*
**Zhaoshu Shi** SHIZHAOSHU@MYHEXIN.COM
*Hithink RoyalFlush Information Network Co.,Ltd, Hangzhou, China*

**Editors:** Berrin Yanıkoğlu and Wray Buntine

## Abstract

Aspect-based-sentiment-analysis (ABSA) is a fine-grained sentiment evaluation task, which analyzes the emotional polarity of the evaluation aspects. Generally, the emotional polarity of an aspect exists in the corresponding opinion expression, whose diversity has great impact on model's performance. To mitigate this problem, we propose a novel and simple counterfactual data augmentation method to generate opinion expressions with reversed sentiment polarity. In particular, the integrated gradients are calculated to locate and mask the opinion expression. Then, a prompt combined with the reverse expression polarity is added to the original text, and a Pre-trained language model (PLM), T5, is finally was employed to predict the masks. The experimental results shows the proposed counterfactual data augmentation method performs better than current augmentation methods on three ABSA datasets, i.e. Laptop, Restaurant, and MAMS.

**Keywords:** ABSA, counterfactual data augment, opinion expression, integrated gradient

## 1. Introduction

Traditional sentiment analysis tasks consider sentences or documents as the object to analyze their sentiment polarity. However, a sentence may contain multiple aspects with different emotional polarities (Xu et al., 2019; Song et al., 2019; Wang et al., 2020; Zhao et al., 2020). This bring much difficulty for the more fine-grained sentiment analysis.

Pre-trained language models (PLMs) that were trained on massive data in an unsupervised manner, possess great natural language understanding ability and have been fine-tuned to solve various downstream tasks effectively, e.g. sentiment analysis, textual entailment, text summarization, question answering, etc. In practice, the samples are generally concatenated with the aspect words as the input of PLMs, and the parameters of PLMs can be freezed or tuned during training (Song et al., 2019). Recently, due to the requirement of few computing resources during fine-tuning and good performance in solving downstream tasks, some new fine-tuned methods, e.g. Prefix tuning (Li and Liang, 2021), P-tuning (Liu et al., 2022), LoRA (Hu et al., 2021) etc., which add some external structures while freezing the parameters of PLMs, have drawn widespread attention.

Aspect-based-sentiment-analysis (ABSA) can be roughly divided into two subtasks: (1) Extracting the evaluation objects from the text. (2) Judgement of the emotional polarity of the objects. Since a sentence might contain multiple aspects and the opinion expressions may exist implicitly in the text, it is hard to identify all the opinion expressions accurately. Meanwhile, the sample diversity in ABSA tasks is usually insufficient, and this further affects the performance of the fine-tuned models.

Being the simplest method, data augmentation which helps to improve the diversity of training samples, can thus be used to alleviate the above issues. Data augmentation methods can be roughly divided into two categories: the modification of existing samples and the generation of new samples (Anaby-Tavor et al., 2019; Kumar et al., 2020), where the modification methods can be further divided into noising (Wei and Zou, 2019), thesauruses (Zhang et al., 2015), machine translation (Sennrich et al., 2015) and language models (Wu et al., 2019; Jiao et al., 2019).

Due to the ability to consider context semantics and alleviate the ambiguity problem, language models are ideal for fine-grained natural language processing (NLP) tasks. However, language model-based methods have shortcomings in limiting the word level and affecting the sentence semantics if there are excessive random substitutions. Meanwhile, as it is a label-preserving method, the modifications are restricted to the same semantic area.

Recently, generative large language models (GLLMs) with billion parameters which are trained on tera-scale tokens can deal with various downstream tasks by one shot, few shot, or even zero shot. The performance of GLLMs is even better than the models fine-tuned on the specific training data of ABSA tasks. However, the high cost of deployment and the strict governmental policy have made the access to GLLMs difficult. Therefore, it is of great significance to develop simple, resource-friendly and effective ABSA method.

This paper proposes a novel counterfactual data augmentation method for ABSA task. The proposed method is a language model-based method and can be mainly divided into two stages. The integrated gradients is first used to identify the opinion words which thereafter will be masked. Next, the prompts with reversed polarity are added to the original sentences and the language model T5 is employed to predict the masks. In this way, the original aspect polarity is reversed while only modifying a few tokens, and therefore, the model can obtain stronger generalization ability. The proposed method was tested on three open datasets, e.g., Restaurant, Laptop and MAMS, and the results show that the proposed method performs better than several common data augmentation methods.

The main contributions of this paper are as follows:

(1) We proposed a two-stage data augmentation method composed of opinion corruption and diverse opinion generation.

(2) The proposed counterfactual data augmentation method is simple and easy to implement in the production environment, and can be combined with other baseline models to further improve the results of ABSA tasks.

The reminder of this paper is arranged as follows. Section 2 introduces the related works. Section 3 presents the proposed counterfactual data augmentation method. The experimental results and conclusion are given in Section 4 and Section 5 respectively.

## 2. Related Work

### 2.1. ABSA

Since the introduction of BERT (Devlin et al., 2018), PLMs based on the Transformer architecture (Vaswani et al., 2017) has transformed the paradigm of NLP. These models are now dominating downstream tasks including ABSA tasks owing to their strong natural language understanding capability. For example, BERT-SPC (Song et al., 2019) leverages PLMs to incorporate both the text and the aspect into the model input, and thus allowing effective modeling of the relationships between aspects and opinion expressions. Furtherly, adversarial training (Karimi et al., 2021), layer aggregation (Karimi et al., 2020) and domain adaption (Rietzler et al., 2019) were integrated to improve the performance.

As a fine-grained NLP task, it is important to incorporate syntactic information into ABSA. This can guide the model to focus on the relevant parts of the aspects, which in turn can help to solve the opinion identification problem more effectively. AdaRNN (Dong et al., 2014) transformed the dependency tree, starting from the aspect words, into a recursive structure which was modeled with Recursive Neutral Network (RNN). As dependency information are not entirely accurate and needs to be corrected for accurate target identification, He et al. (2018) incorporated syntactic information with attention mechanism. Zhang et al. (2019) then applied proximity weight on both position and dependency, Wang et al. (2020) then reshaped and pruned the original dependency tree. Yet the derived dependency tree can only represent external criteria, conflicting with the knowledge of the fine-tuned PLMs. Dai et al. (2021) proposed a distance-based method which derivied the dependency trees from fine-tuned PLMs.

Another challenge is the multi-aspect problem, where features of different aspects would affect each other. Liang et al. (2021); Wang et al. (2022) employed contrastive training objects to tackle the challenge. Yang and Li (2021) designed a local sentiment aggregation method that can facilitate mutual learning among aspects, enabling the discovery of implicit expressions.

In addition, with the development of GLLM, auto-regressive models have been applied for ABSA tasks. For example, Mao et al. (2021) converted ABSA to a text-to-text task, marking the required emotional elements in the original sentence and using it as the target sequence for the generative model to learn the mapping relationship. Scaria et al. (2023) directly used the instruction-tuning based model and achieved good results.

### 2.2. Data Augmentation

Data augmentation refers to the modification and expansion of the original data, therefore introducing more samples and increasing the diversity of the training set. Modification means to modify the elements in the sentence without destroying the original sentence structure and label. EDA (Wei and Zou, 2019) is a simple modification method consisting of synonym replacement, random deletion and insertion. CBERT (Wu et al., 2019) proposed a mask and predict mechanism where the words were masked randomly and then predicted by BERT. BackTranslation (Sennrich et al., 2015) that can get high-quality, richly diverse samples by translating back and forth, also draw widespread attention. However, the flip side of the coin is that modifications are restricted to the same semantics as a label-preserving method. Expansion means creating new data by generation methods and sampling from it. It is a more flexible way and can be designed for different task descriptions. LAMBDA (Anaby-Tavor et al., 2019) transforms the classification dataset into a seq2seq dataset and fine-tuned it on GPT-2, then generates new sentences with specific labels. However, such process requires extensive computing resources.

There are also some proprietary augmentation methods for ABSA task. Chen et al. (2022) extended unsupervised data augmentation methods to span-level. Zhang et al. (2022) introduced datasets from sentiment analysis and merged them using pseudo-labels. Based on supervised attention, Liang et al. (2021) extracted crucial information and utilized it to generate augmented data. Wang et al. (2022) applied contrastive learning to distinguish different polarities, where negative samples are generated by T5 fine-tuned with Prompt Tuning. Hsu et al. (2021) introduced a two-stage method composed of selective perturbed masking and label-preserving token replacement.

### 2.3. Interpretability

Interpretability refers to the extent to which a person can comprehend the rationale behind a decision. There are two mainstreaming solutions for interpretability: attention mechanism (Bahdanau et al., 2014) based methods and saliency methods. Notwithstanding that attention brings transparency to the model by attention weights between two input units, someone believe that it is not sufficiently interpretable. Pruthi et al. (2019) proposed a deceptive self-attention method which help explain the interaction of information in transformer. Jain and Wallace (2019) discovered that the adversarial attention weights lead to the same prediction as the original ones. Serrano and Smith (2019) found that model did not recognize the most important expressions by intermediate representation erasure. To argue this, Wiegreffe and Pinter (2019) believed that attention can be used as explanation in some scenarios, and verified through experiments.

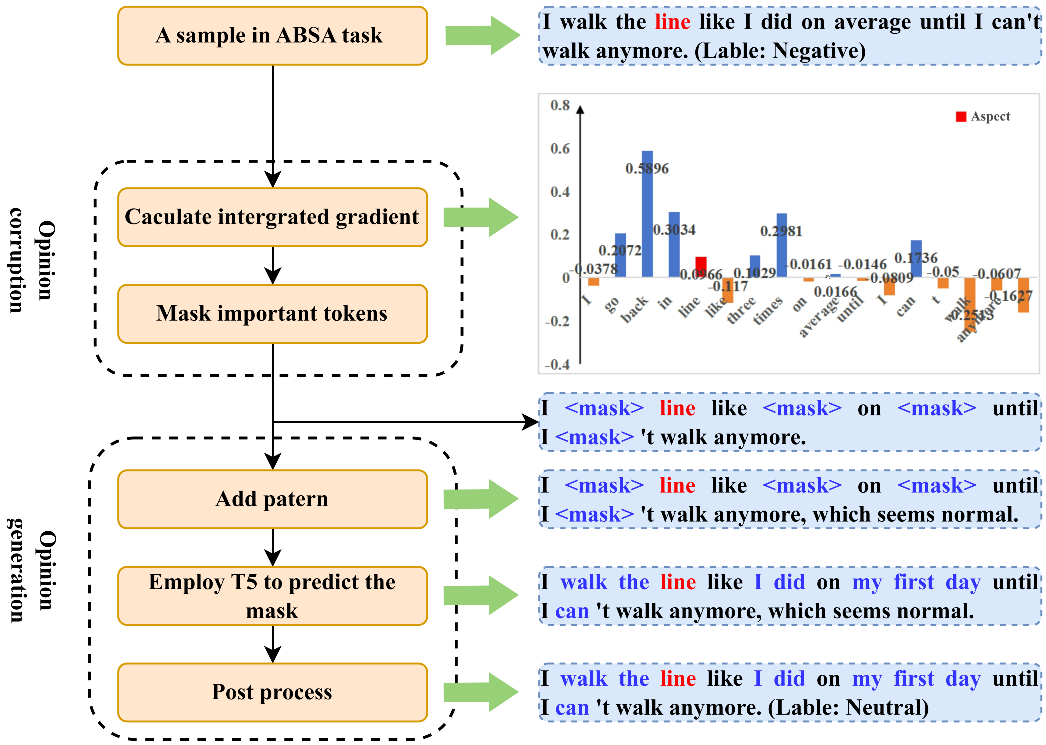
Saliency methods can be further divided into erasure-based and gradient-based methods. Zeiler and Fergus (2014); Li et al. (2016) pointed out the importance of input units by erasing them at the input level and dimension level. Li et al. (2015) used the absolute value of the gradient to measure the sensitivity of the input to the label in the ABSA task, while Denil et al. (2014) used the product of the gradient and the input as a measure. Sundararajan et al. (2017) first proposed three axioms, namely sensitivity, implementation invariance and completeness.

## 3. Method

This section presents the details of the proposed counterfactual data augment method for ABSA.

### 3.1. Framework

As shown in Fig.1, the proposed counterfactual data augmentation method for ABSA is mainly composed of two steps, i.e. opinion corruption and opinion generation. The opinion corruption step is designed to identify and mask the most important tokens to the target label. As for the opinion generation step, the designed prompt is added to the masked



I walk the line like I did on average until I can't walk anymore. (Lable: Negative)

I <mask> line like <mask> on <mask> until I <mask> 't walk anymore.

I <mask> line like <mask> on <mask> until I <mask> 't walk anymore, which seems normal.

I walk the line like I did on my first day until I can 't walk anymore, which seems normal.

I walk the line like I did on my first day until I can 't walk anymore. (Lable: Neutral)

Figure 1: Framework of the proposed counterfactual data augmentation framework.

sentences and the masked tokens will be predicted by T5. Finally, the added prompt is removed from the new sentence and the new label is obtained by a post process.

### 3.2. Opinion corruption

To identify the associated opinion words of the given aspect, a quantitative analysis of opinion words is implemented by integrated gradients as described in Sundararajan et al. (2017).

Assuming the original dataset, training set and test set are $D$, $D_{train}$, $D_{test}$ respectively. $x_i$, $A_i$, $y_i$ represent a training sample, the aspects and the corresponding label respectively. The aspects may contain $k$ aspects $A_i = \{a_{i0}, ..., a_{ik}\}$.

The first step is to train a classifier $M_{base}$ in advance to calculate integrated gradients. Since the existence of data imbalance problem, the balanced cross entropy(BCE) thereby is introduced in the loss function. Suppose in one batch, the numbers of three labels are $n_0, n_1, n_2$, and the standard cross entropy loss is $L(M(x), y)$, so the BCE can be described by equation 1:

$$L_{balanced} = \sum_{i=0}^{2} (1 - \frac{n_i}{\sum_i n_i}) L(M(x), y) \tag{1}$$

Then the contribution of each token in $x_i = \{r_0, r_1, ..., a_0, ..., r_l, [SEP], a_0\}$ to the label $y_i$ is calculated. $r_i$ denotes token, $a_0$ denotes aspect.

Specifically, the integrated gradients are the path integral of gradient from a baseline $x_{i0}$ to the input $x_i$. Here, we replace all tokens except ones in aspects set with [PAD] as the baseline. Then, the integrated gradients can be represented as equation 3.

$$x_{i0} = \{[PAD], [PAD], ..., a_0, ..., [PAD], [SEP], a_0\} \tag{2}$$

$$ig(x_i) = (x_i - x_{i0}) \times \int_{\alpha=0}^{1} \frac{\partial M(x_{i0} + \alpha(x_i - x_{i0}))}{\partial x_i} d\alpha \tag{3}$$

where $ig(x_i)$ is the attribution of each token in $x_i$ to the label $y_i$.

In practice, it is not possible to calculate the consecutive integrals, and thereby we obtain the integrated gradients by a linear interpolation operation. The interpolated sample $\overline{x_i}$ is as shown in equation 4.

$$\overline{x_i} = \{x_{i0}, x_{i1}, ..., a_{iS}, x_i\} \tag{4}$$

$$x_{ij} = x_{i0} + j\frac{x_i - xi0}{S} \tag{5}$$

where $S$ is the interpolation number.

The input $\{\overline{x_i}, y_i\}$ is then passed into the model $M_{base}$, and a forward and a backward operation are performed to obtain the attribution of each token as shown in following equation.

$$attr(x_i) = Norm_{emb}(\sum_j grad(x_{ij})) \tag{6}$$

Here, the large value of $attr(x_i)$ indicates the higher contribution of $x_i$, and vice versa.

Next, the tokens with higher value than the threshold $thr_{con}$ are masked and the continuous masks will be merged, where $thr_{con} = topK(attr(x_i, floor(\frac{len(attr(x_i))}{3})))$. For instance, in a raw training sentence "Maximum sound isn't nearly as loud as it should be [SEP] Maximum sound", the aspect words are "Maximum sound". The tokens satisfied the threshold are "isn", "nearly", "as", "loud", "be", so the masked sample is "Maximum sound $\langle mask \rangle$ 't $\langle mask \rangle$ as it should $\langle mask \rangle$".

### 3.3. Opinion generation

In opinion generation step, the artificial prompts are first added to the obtained corrupted samples. Compared with the soft prompts, hard prompts methods require no fine-tuning making them more efficient and interpretable.

Assuming the masked sample and it's corresponding label are $\acute{x}_i$ and $\acute{y}_i$, the new sample with added prompts and the new label are $\tilde{x}_i$ and $\tilde{y}_i$ respectively. Consider the above example, the original label negative. Therefore, we should add a positive or neutral prompt, for example "Maximum sound $\langle mask \rangle$ 't $\langle mask \rangle$ as it should $\langle mask \rangle$, which is great!".

Then, the new sample is passed into T5 model to predict the mask tokens and the generation sample $\overline{x_i} = T5(\tilde{x}_i)$ is obtained. Moreover, by removing the added prompt

$pattern_i$ in $\overline{x_i}$, we can acquire the final generated sample $\hat{x_i}$. In the above example, the final augmented sample is “Maximum sound quality ’thumping’ as it should be”.

However, the sentiment polarities of counterfactual samples may be shifted from $\acute{y_i}$, thereby needing assistance from baseline model $M_{base}$ to determine the final label as follows:

$$\tilde{y_i} = \begin{cases} \tilde{y_i} & \text{if } argmax(M_{base}(\hat{x_i})) = \tilde{y_i} \text{ and } max(M_{base}(\hat{x_i})) > thr_{con} \\ argmax(M_{base}(\hat{x_i})) & \text{else} \end{cases} \tag{7}$$

where $thr_{con}$ is the probability threshold. It should note that we add several homogeneous for each reversed label and choose the label with maximum probability fluctuation compared with original sample as the final label. Finally, the augmented data are merged with the original training set.

## 4. Experiment

### 4.1. Experimental Settings

#### 4.1.1. Datasets

The proposed counterfactual data augmentation method was tested on three common datasets, i.e., SemEval 2014 Restaurant, Laptop (Pontiki et al., 2014) and MAMS (Jiang et al., 2019). The statistics of the datasets are shown in Table 1. Following previous research (Wang et al., 2022; Hsu et al., 2021), we adopt accuracy and Macro-F1 as the metrics to evaluate the performance. Note that the original datasets do not include validation set.

Table 1: The statistics of the datasets.

| Dataset | Positive | | Neutral | | Negative | |
|---|---|---|---|---|---|---|
| | Train | Test | Train | Test | Train | Test |
| Laptop | 994 | 341 | 870 | 128 | 464 | 169 |
| Restaurant | 2164 | 728 | 807 | 196 | 637 | 196 |
| MAMS | 3379 | 400 | 2763 | 329 | 5039 | 607 |

#### 4.1.2. Comparison experiment settings

First, we compare the proposed counterfactual data augmentation method with other data augmentation methods as shown in the following.

- **EDA** (Wei and Zou, 2019): A simple augmentation method including random insertion, deletion, replacement.
- **BackTranslation** (Sennrich et al., 2015): Translate the text into another language by machine translation models and then translate it back into the original language. Here, the experimental samples are translated into Chinese which are then translated back to English.

- **C$^3$DA** (Wang et al., 2022): A cross-channel data augmentation method aiming to generation negative samples for contrastive learning.
- **Senti-SPM** (Hsu et al., 2021): A method composed of selective perturbed masking (SPM) and label-preserving token replacement.

In addition, the BCE was chosen as the loss function of ABSA to be consistent with C$^3$DA, and the foundation model is Roberta-base.

Second, the results obtained by different baseline methods of ABSA, i.e., ASGCN (Sun et al., 2019), PWCN (Zhang et al., 2019), RGAT (Wang et al., 2020), SPC (Song et al., 2019), MLP (Dai et al., 2021), AEN (Song et al., 2019), LCF (Yang et al., 2019), were compared when using different foundation models, e.g., Bert, Roberta and glove.840B.300d. Here, the general cross entropy is selected as the loss function of ABSA to make a fair comparison.

#### 4.1.3. Hyper Parameters

Following previous studies, the batch size is set as 32, the learning rate is 2e-5 for Bert and Roberta and 1e-3 for glove.840B.300d. Three random seeds was used in all the experiments and we present the average results.

### 4.2. Augmentation Comparison Result

To evaluate the effectiveness of the proposed counterfactual data augmentation method, we compared it with other data augmentation methods and the results are presented in Table 2.

Table 2: Experimental results of ABSA tasks using different data augmentation methods.

| Method | Laptop | | Restaurant | | MAMS | |
|---|---|---|---|---|---|---|
| | Acc | F1 | Acc | F1 | Acc | F1 |
| RoBERTa-SPC | 76.91 | 70.47 | 84.73 | 76.15 | 83.61 | 82.88 |
| C$^3$DA | 81.83 | 78.46 | 87.11 | 81.63 | - | - |
| EDA | 83.07 | 80.22 | 88.21 | 83.04 | 84.95 | 84.49 |
| BackTranslation | 82.6 | 79.13 | 88.12 | 82.36 | 84.73 | 84.45 |
| Senti-SPM & Seq2Seq | 83.7 | 80.82 | 88.39 | 83.05 | - | - |
| Counterfactual | **83.86** | **81.39** | **89.2** | **84.14** | **85.33** | **84.87** |

From Table 2, it can be seen that data augmentation based methods can obviously improve the accuracy and F1 score on Laptop and Restaurant datasets compared with the baseline method RoBERTa-SPC, while obtain slight improvement on MAMS datasets that contain adequate training samples. EDA outperforms BackTranslation on all three datasets. This may be because EDA causes less damage to the original text as EDA employs random replacement, deletion and insertion to introduce diverse opinion expressions, while BackTranslation may destroy the relationship between emotional expressions and aspect words.

It is also can be seen that Senti-SPM&Seq2Seq achieve the best results except our proposed counterfactual data augmentation method. However, Senti-SPM&Seq2Seq just replace a few unimportant tokens ignoring some important tokens that containing rich semantics.

As for our proposed counterfactual data augmentation method, it achieves the best results compared with the other four data augmentation methods. In addition, the proposed method performs good robustness and generalization as the counterfactual operation enrich the diversity of the original samples.

Furthermore, the effects of different foundation models, i.e., BERT, RoBERTa and Static embedings, and different baseline methods, i.e., ASGCN, PWCN, RGAT, SPC, AEN, MLP and LCF, were evaluated and the results are shown in Table 3.

Table 3: Results of the proposed counterfactual data augmentation method combined with different foundation models and baseline methods.

| Models | | Restaurant | | Laptop | |
|---|---|---|---|---|---|
| | | Acc | F1 | Acc | F1 |
| Static Embeddings | ASGCN | 80.09 | 71.01 | 74.27 | **69.98** |
| | + Counterfactual | **80.77** | **71.08** | **74.33** | 69.72 |
| | PWCN | 81.10 | 72.29 | 75.34 | 71.21 |
| | + Counterfactual | **82.26** | **74.16** | **76.02** | **71.86** |
| | RGAT | 81.19 | 71.7 | 73.09 | 68.37 |
| | + Counterfactual | **82.00** | **73.16** | **74.61** | **69.64** |
| BERT | BERT-SPC | 84.02 | 74.94 | 76.42 | 72.06 |
| | + Counterfactual | **84.73** | **78.03** | **77.59** | **72.49** |
| | RGAT-BERT | 85.03 | **78.01** | 78.69 | 74.31 |
| | + Counterfactual | **85.21** | 77.77 | **79.05** | **75.36** |
| | BERT-MLP | 84.33 | 77.22 | 77.77 | 73.39 |
| | + Counterfactual | **84.51** | **77.59** | **78.14** | **73.96** |
| | AEN-BERT | 81.30 | 71.13 | 77.43 | 72.02 |
| | + Counterfactual | **81.30** | **71.73** | **77.90** | **72.78** |
| | LCF-BERT | 85.27 | 78.81 | 77.83 | 73.23 |
| | + Counterfactual | **85.40** | **79.69** | **78.53** | **74.59** |
| RoBERTa | RoBERTa-SPC | 84.73 | 76.15 | 76.91 | 70.47 |
| | + Counterfactual | **86.67** | **79.73** | **77.79** | **72.93** |
| | RGAT-RoBERTa | 85.3 | 77.75 | 77.64 | 73.9 |
| | + Counterfactual | **86.07** | **79.63** | **78.68** | **74.87** |
| | RoBERTa-MLP | 86.79 | 79.86 | 83.86 | 80.41 |
| | + Counterfactual | **86.83** | **81.03** | **83.86** | **80.61** |

It can be seen from Table 3 that whatever the foundation models, the proposed counterfactual data augmentation method can improve the accuracy and F1 score compared with baseline methods. Since RoBERTa model was trained on larger scale data and can obtain better text embedding, combined the proposed counterfactual data augmentation method with baseline methods acquire the best results. Meanwhile, the results also show that the proposed counterfactual data augmentation method can be easily combined with baseline methods in production.

### 4.3. Ablation Study

In this section, we further investigate the impact of different masking and prompting strategies on the results as shown in Table 4.

- **Ramdom-Mask**: Replace the integrated gradients based mask strategy with random mask.
- **Label-Preserve**: Add prompt with the same polarity of the original sample during opinion generation.

Table 4: Results of different mask strategy and prompting method.

| Method | Laptop | | Restaurant | | MAMS | |
|---|---|---|---|---|---|---|
| | Acc | F1 | Acc | F1 | Acc | F1 |
| RoBERTa-SPC | 76.91 | 70.47 | 84.73 | 76.15 | 83.61 | 82.88 |
| Random-Mask | 82.29 | 79.1 | 88.04 | 82.58 | 85.18 | 84.85 |
| Label-Preserve | 83.7 | 80.82 | 88.39 | 83.05 | 84.96 | 84.43 |
| Counterfactual | **83.86** | **81.39** | **89.2** | **84.14** | **85.33** | **84.87** |

From Table 4, it can be seen that the proposed counterfactual data augmentation method which employ an integrated gradient-based mask strategy, performs better than random mask strategy. It is because that random mask may create samples that retain the same emotional semantics, and thus not increase the semantic diversity significantly. Since label-preserve prompting is likely to generate synonyms of original tokens, the semantic diversity of samples is not enriched. Therefore, it can be found that counterfactual prompting method obtain better accuracy and F1 score on all three experimental datasets.

### 4.4. Augmented sample analysis

Here, an analysis of augmented samples by different methods is presented to provide a more comprehensive comparison as shown in Table 5.

From Table 5, it can be seen that EDA will randomly replaced or deleted some tokens, may leading the change of emotional semantics. BackTranslation maintain the semantics of original sample, but may modify the aspect words. Random-Mask just mask some tokens randomly and thus fail to identify opinion expressions. However, the proposed counterfactual data augmentation method make the key opinion expression changed while modify the emotional semantics, and thus increase the diversity of samples.

## 5. Conclusion

This paper proposed a novel and simple counterfactual data augmentation method for ABSA. An integrated gradient-based method is used to identify key opinion expressions which are masked and then will be predicted by T5 to obtain rich opinion expressions. The

Table 5: Case study. The two cases come from the Restaurant and MAMS datasets respectively. The first one is only a single aspect word, and the second one has three aspect words, which are enhanced based on ”portions”.

| Methods | Examples |
|---|---|
| Source | I go back in **line** like three times on average until I can’t walk anymore. |
| EDA | like go back in **line** i three times on average until i cant walk anymore. |
| BackTranslation | I have to requeue an average of 3 times until I can no longer walk. |
| Random-Mask | I go ( in **line** like three times) on average until I can ’t walk anymore. |
| Counterfactual | I walk the **line** like I did on my first day until I can ’t walk again. |
| Source | The food is right out of heaven, arrive hungry because the **portions** are huge but not the prices. |
| EDA | the food is right out of heaven arrive hungry come because the **portions** are huge merely but not the prices |
| BackTranslation | The food was heaven, arrived hungry as the **portions** were huge but not overpriced. |
| Random-Mask | The food were so right out of heaven, arrive hungry because are **portions** is just too huge but not the prices. |
| Counterfactual | the food was out of order when we arrive, **portions** were small but not at reasonable food and prices. |

experiments show that the proposed counterfactual data augmentation method is superior to other augmentation methods, and achieved good results on public datasets. It is expected that the proposed counterfactual data augment method will have the opportunity to expand to other fine-grained NLP tasks in the future.

## References

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. Not enough data? deep learning to the rescue! *CoRR*, abs/1911.03118, 2019. URL http://arxiv.org/abs/1911.03118.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

David Z Chen, Adam Faulkner, and Sahil Badyal. Unsupervised data augmentation for aspect based sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6746–6751, 2022.

Junqi Dai, Hang Yan, Tianxiang Sun, Pengfei Liu, and Xipeng Qiu. Does syntax matter? a strong baseline for aspect-based sentiment analysis with roberta. *arXiv preprint arXiv:2104.04986*, 2021.

Misha Denil, Alban Demiraj, and Nando De Freitas. Extraction of salient sentences from labelled documents. *arXiv preprint arXiv:1412.6815*, 2014.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 49–54, 2014.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. Effective attention modeling for aspect-level sentiment classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1121–1131, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1096.

Ting-Wei Hsu, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. Semantics-preserved data augmentation for aspect-based sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4417–4422, 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. A challenge dataset and effective models for aspect-based sentiment analysis. 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 6280–6285, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1654. URL https://aclanthology.org/D19-1654.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling BERT for natural language understanding. *CoRR*, abs/1909.10351, 2019. URL http://arxiv.org/abs/1909.10351.

Akbar Karimi, Leonardo Rossi, and Andrea Prati. Improving bert performance for aspect-based sentiment analysis. *arXiv preprint arXiv:2010.11731*, 2020.

Akbar Karimi, Leonardo Rossi, and Andrea Prati. Adversarial training for aspect-based sentiment analysis with bert. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8797–8803. IEEE, 2021.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. Data augmentation using pre-trained transformer models. *CoRR*, abs/2003.02245, 2020. URL https://arxiv.org/abs/2003.02245.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. Visualizing and understanding neural models in nlp. *arXiv preprint arXiv:1506.01066*, 2015.

Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*, 2016.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353.

Bin Liang, Wangda Luo, Xiang Li, Lin Gui, Min Yang, Xiaoqi Yu, and Ruifeng Xu. Enhancing aspect-based sentiment analysis with supervised contrastive learning. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 3242–3247, 2021.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, 2022.

Yue Mao, Yi Shen, Chao Yu, and Longjun Cai. A joint training dual-mrc framework for aspect based sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13543–13551, 2021.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland, August 2014. Association for Computational Linguistics. doi: 10.3115/v1/S14-2004. URL https://aclanthology.org/S14-2004.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. Learning to deceive with attention-based explanations. *arXiv preprint arXiv:1909.07913*, 2019.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification. *arXiv preprint arXiv:1908.11860*, 2019.

Kevin Scaria, Himanshu Gupta, Saurabh Arjun Sawant, Swaroop Mishra, and Chitta Baral. Instructabsa: Instruction learning for aspect based sentiment analysis. *arXiv preprint arXiv:2302.08624*, 2023.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*, 2015.

Sofia Serrano and Noah A Smith. Is attention interpretable? *arXiv preprint arXiv:1906.03731*, 2019.

Youwei Song, Jiahai Wang, Tao Jiang, Zhiyue Liu, and Yanghui Rao. Attentional encoder network for targeted sentiment classification. *arXiv preprint arXiv:1902.09314*, 2019.

Kai Sun, Richong Zhang, Samuel Mensah, Yongyi Mao, and Xudong Liu. Aspect-level sentiment analysis via convolution over dependency tree. 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 5679–5688, 2019.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Bing Wang, Liang Ding, Qihuang Zhong, Ximing Li, and Dacheng Tao. A contrastive cross-channel data augmentation framework for aspect-based sentiment analysis. *arXiv preprint arXiv:2204.07832*, 2022.

Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang. Relational graph attention network for aspect-based sentiment analysis. *arXiv preprint arXiv:2004.12362*, 2020.

Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*, 2019.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. Conditional bert contextual augmentation. In *Computational Science–ICCS 2019: 19th International Conference, Faro, Portugal, June 12–14, 2019, Proceedings, Part IV 19*, pages 84–95. Springer, 2019.

Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. Bert post-training for review reading comprehension and aspect-based sentiment analysis. *arXiv preprint arXiv:1904.02232*, 2019.

Heng Yang and Ke Li. Improving implicit sentiment learning via local sentiment aggregation. *arXiv e-prints*, pages arXiv–2110, 2021.

Heng Yang, Biqing Zeng, Jianhao Yang, Youwei Song, and Ruyang Xu. A multi-task learning model for chinese-oriented aspect polarity classification and aspect term extraction. *CoRR*, abs/1912.07976, 2019. URL http://arxiv.org/abs/1912.07976.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.

Chen Zhang, Qiuchi Li, and Dawei Song. Syntax-aware aspect-level sentiment classification with proximity-weighted convolution network. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 1145–1148, 2019.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *CoRR*, abs/1509.01626, 2015. URL http://arxiv.org/abs/1509.01626.

Yiming Zhang, Min Zhang, Sai Wu, and Junbo Zhao. Towards unifying the label space for aspect-and sentence-based sentiment analysis. *arXiv preprint arXiv:2203.07090*, 2022.

Pinlong Zhao, Linlin Hou, and Ou Wu. Modeling sentiment dependencies with graph convolutional networks for aspect-level sentiment classification. *Knowledge-Based Systems*, 193:105443, 2020.