International Journal of Machine Learning and Cybernetics
https://doi.org/10.1007/s13042-024-02180-w

ORIGINAL ARTICLE

Prompt-based data labeling method for aspect based sentiment analysis

Kun Bu[1] · Yuanchao Liu[1]

Received: 11 July 2023 / Accepted: 10 April 2024

Abstract
ABSA aims to extract aspect terms and corresponding sentiment from unstructured texts. Supervised approaches are widely used in existing ABSA models because of their model maturity, and most of them usually need large-scale training data to deal with over-fitting. However, in real scenarios, the labeled data is difficult to obtain, thus the performance is adversely influenced. To address these issues, this paper proposes a prompt-based data augmentation method, enabling it to overcome small data problems by expanding the sample size in the training corpus. Our approach computes the relationship between the prompt templates and unlabeled data and then assigns labels to expand the training data. To achieve this, we formulate it as a data filtering problem and implement it with Natural Language Inference models. The experimental results on four well-studied datasets demonstrate that our model not only achieves results on par with existing state-of-the-art data augmentation methods on a few occasions but also significantly improves the effectiveness of existing ABSA models on most occasions, indicating its strong robustness in various base ABSA models. Further discussion shows that prompt learning can help the model mark data from unlabeled datasets, which explains its effectiveness in data augmentation.

Keywords Natural language processing · Data augmentation · Prompt learning · Aspect based sentiment analysis · Neural network

1 Introduction

As shown in Fig.1, Aspect Based Sentiment Analysis (ABSA) aims to identify the sentiment polarity of one or more specific aspects [1]. The input is a review and the output is the aspects and their corresponding sentiment labels (Pos, Neu, Neg) in this sentence. According to Bu et al. [1], as training data volume increases, the performance of document-level sentiment analysis shows a downward trend, while ABSA shows an upward trend.

ABSA is a highly data-dependent task in Natural Language Processing (NLP), it has derived many sub-tasks, making ABSA one of the most challenging tasks that have attracted the attention of a large number of researchers. Due to the complexity of ABSA, the models require a large amount of labeled training data. Currently, many state-of-the-art (SOTA) ABSA methods rely on supervised learning due to their high rates of effectiveness [2]. In this paper, we design reasonable and low-compute resource augmentation strategies for ABSA. Figure 2 shows an example. The unlabeled review contains one or more aspects and corresponding sentiment that can serve as the training data for the ABSA model. Our prompt-based model can locate the aspects, recognize the corresponding polarities, and extract them as the data label.

Data Augmentation (DA) was initially applied in computer vision [3, 4] and had been widely applied in NLP [5, 6]. Common methods include random addition, deletion, and exchange; The Synonym Replacement (SR) is a simple and intuitive data augmentation method, which replaces some words in the source sentence with synonyms in WordNet or similar words in the word vector [7]. Another method is back-translation. It needs to translate the original sentence between the source language and multiple target languages many times. Then translate it back to the source language [8]. Text generation-based DA model is also a popular method. Generative Adversarial Network

✉ Kun Bu
21b903084@stu.hit.edu.cn

[1] School of Computer Science and Technology, Harbin Institute of Technology, No.92, West Da Zhi Street, Nangang District, Harbin 150001, Heilongjiang, People's Republic of China

Published online: 23 May 2024

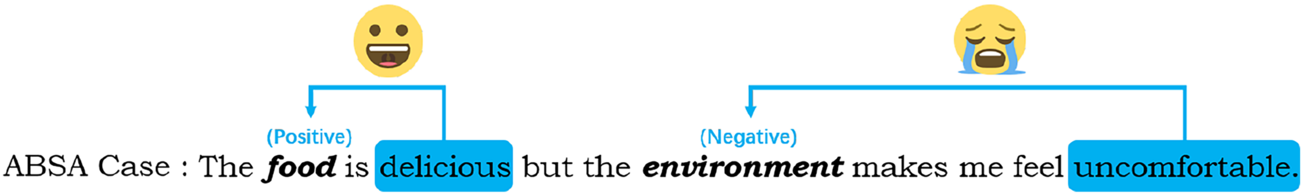
(Positive)
(Negative)
ABSA Case : The *food* is delicious but the *environment* makes me feel uncomfortable.

Fig. 1 The main purpose of ABSA

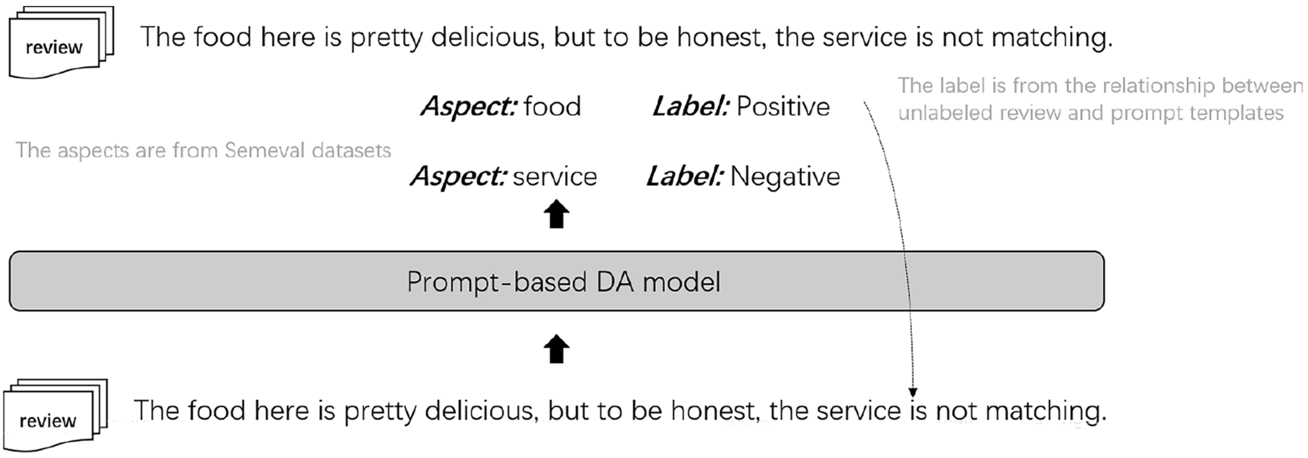
review
The food here is pretty delicious, but to be honest, the service is not matching.
Aspect: food
Label: Positive
The label is from the relationship between unlabeled review and prompt templates
The aspects are from Semeval datasets
Aspect: service
Label: Negative
Prompt-based DA model
review
The food here is pretty delicious, but to be honest, the service is not matching.

Fig. 2 Our DA system takes a unlabeled data as the input and extracts the aspects and corresponding label from the unlabeled as the output

(GAN) [9] and Variational AutoEncoder (VAE) [10] are two generation models based on neural networks, the text is generated based on the input text and can be used for data augmentation of sentiment analysis [11]. Recently, the widespread use of pre-trained language models has also brought new methods to DA [12]. Kobayashi [13] proposed to randomly replace words with words predicted by the pre-trained language models. Existing methods have effectively improved the performance of the ABSA model. Wang et al. [14] proposed a contrastive cross-channel data augmentation framework to generate more domain and multi-dimensional samples, and trained a more robust ABSA model based on these generated data. However, none of the methods is perfect, because semantic distortion, syntax errors, and other data noises may occur in the augmentation process.

Small changes in training data can mislead the model into making incorrect predictions [15]. Compared with working with existing data, we believe that automated labeling of true unlabeled data is a more worthwhile direction to explore. Li et al. [16] summarized the common DA methods in NLP. They also believed that external datasets have very high data values. Compared with consuming a lot of resources to work on existing datasets, automatic annotation of external datasets is more meaningful, but data annotation is an arduous task.

Data is the foundation of Artificial Intelligence (AI). However, most existing ABSA datasets are manually annotated, which has low efficiency and cannot guarantee the accuracy of annotated data. The uncertainty of manual data annotation can affect the distribution of data in the datasets. The proportion of various types of data in the datasets will directly affect the performance of various ABSA models [17–19]. With the development of neural networks, the structure of neural networks becomes more and more complex [20, 21]. Correspondingly, the demand for data rises dramatically. The problem of insufficient data is alleviated by pre-training models that are pre-trained on unlabeled data and then fine-tuned in downstream tasks. However, in the fine-tuning phase, insufficient training data can lead to over-fitting. The model performance is negatively affected by the lack of sufficiently labeled data. The prompt-based DA method has been widely applied in other NLP domains [22–25], but it is rare in ABSA. We need an automatic data labeling method to provide more high-quality labeled data.

According to Huang et al. [26], in Amazon and Yelp (document-level sentiment dataset), there is a large amount of data that can be used as ABSA training data. However, in the document-level sensitive dataset, the data is only divided into 1 to 5 ratings, which requires further processing.

In this paper, we propose a prompt-based DA method. The contributions of our paper are summarized as follows:

(1) To our knowledge, the data augmentation method of ABSA via prompt learning is rare.
(2) We propose a prompt-based data augmentation method for ABSA and sub-tasks, which provides more real and diverse data with the help of external knowledge. Unlike previous generative data augmentation methods, our model only requires low computational resources.
(3) We provide qualitative analysis and discussions as to why our augmentation method works and test its implementation.

The structure of this paper is as follows: The background and related work, which includes a discussion about existing ABSA and DA models, and the development of prompt learning is introduced in Sect. 2. Our method and the details of the implementation of automated data annotation are introduced in Sect. 3. Then we present the experiment settings, results, discussion, case study, and ablation study in Sect. 4. Finally, we give the conclusion and future work in Sect. 5

2 Related work

2.1 Aspect-based sentiment analysis

ABSA is a classic task. It has great significance for practical application. Zhao et al. [27] regarded this problem as the joint extraction of terms and relationships, and designed a span-based multi-task learning (SpanMlt) framework to jointly extract aspects/views and pairing relationships. Chen et al. [28] proposed a model containing two channels to extract aspect/viewpoint terms and relationships respectively. Two synchronization mechanisms are further designed to realize the information interaction between the two channels. End-to-End ABSA (E2E-ABSA) is used to extract the aspect term and its corresponding sentiment polarity at the same time. It can be divided into two subtasks. Some common ideas frequently appear in different models, the Relationship Aware Collaborative Learning (RACL) framework proposed by Chen and Qian [29] explicitly models the interaction of multiple tasks, and uses the relationship propagation mechanism to coordinate these tasks. Liang et al. [30] further designed a routing algorithm to improve knowledge transfer between these tasks. One of the subtasks is Aspect Category Sentiment Analysis (ACSA). The aspects extracted by E2E-ABSA must be clear in the sentence, while ACSA can be extracted whether implicit or explicit. Because of this feature, ACSA is more widely used in industry. Cai et al. [31] proposed a hierarchical classification method to solve the ACSA problem: Hier-GCN first recognizes aspect categories, then jointly predicts the sentiment of each recognition category. Similarly, Li et al. [32] used the shared sentiment prediction layer to share sentimental knowledge between different categories to alleviate the problem of insufficient data. Liu et al. [33] used the Seq2Seq modeling paradigm to solve the ACSA problem. Based on the pre-trained generation model, they use natural language sentences to represent the required output, and its performance is better than previous models.

In the real world, comment texts are mostly informal expressions with complex grammatical structures. Li et al. [34] proposed DualGCN, which combines syntactic structure complementarity and semantic correlation. The SynGCN module (with rich syntactic knowledge) aims to reduce dependency analysis errors, while the SemGCN module (self-attention mechanism) aims to capture semantic correlations. Zhong et al. [35] introduced external knowledge in the process of solving the ABSA problem, and through complementary information between external knowledge and context (combining external knowledge with context and syntax), captured emotional features from three perspectives: context-based, syntax-based, and knowledge-based. Context and syntax extract features through pre-trained word embedding representations. Specifically, the context is encoded through BiLSTM; The process of syntax encoding is to first establish syntactic dependencies, obtain the adjacency matrix in the sentence, and then encode it through two layers of GCN. Introduce wordnet's knowledge graph as external knowledge and learn knowledge embedding through semantic matching methods. The knowledge representation of specific aspects is learned through the soft attention mechanism.

Until now, one of the main challenges for ABSA tasks lies in the lack of labeled data, especially with the widespread application of large-scale pre-trained models, many DA methods are no longer able to improve model performance [36].

2.2 Data augmentation and prompt learning

The increase in training data does not have a simple linear relationship with model performance, but it cannot be denied that the amount of data is still important for AI models. Here are some examples, for computer vision, the RGB channels' rotations and changes; For speech recognition, the change of sound and speed etc. [12]. DA models should contain source data D, an algorithm A, and new data N. The early ideas can be found in LeNET [37]. Adding noise to existing data is a common method. Coulombe et al. [38] added spelling errors that are common in daily life to the data, resulting in an additional 1.5% increase in XGBoost. Similar studies also include [39, 40] etc.

Embedding replacement is also a popular method. Wang and Yang [41] used KNN to embed the training data words' substitution with the best effect. Compared to the baseline, their logistic regression-based method achieved 2.4% F1 improvement. Similar studies also include [42–44] etc. Their difference lies in the choice of embedded-based words.

Pre-trained language models are widely used in all domains of NLP, and DA is no exception. Wu et al. [45] improved the structure of BERT (c-BERT) through labeled conditional methods, effectively enhancing baseline performance on multiple tasks. However, the performance of c-BERT is not good enough under low data resources. To address this drawback, Hu et al. [46] introduced reinforcement learning on the basis of c-BERT, further improving its model performance. The related work is also reflected in the work from Qu et al. [47] and Anaby Tavor et al. [48].

Prompt learning can effectively improve the performance of pre-trained models [49]. Brown et al. suggested that large-scale models can greatly exploit their inference and understanding capabilities with the help of suitable templates. However, it is initially designed for large pre-trained models. Researchers tried to use it on more general models such as ALBERT etc. Schutze et al. [50] proposed PET (Pattern-Exploiting Training) by text classification tasks and tried to convert all classification tasks to completion blanks consistent with the Masked Language Model (MLM), they designed the important components of Prompt-Tuning, including Pattern (Template) and Verbalizer.

We denote Template as $\mathcal{T}$ and denote Verbalizer as $\mathcal{V}$. The above two components are uniformly defined as Pattern-Verbalizer-Pair (PVP) in this work. Their formal description is shown in Eq. 1.

$$p(y \mid x) = \prod_{j=1}^{n} p\left([\text{mask}]_j = \mathcal{V}(y) \mid \mathcal{T}(x)\right) \tag{1}$$

where x is the sequence, y is the corresponding label.

Figure 3 shows an example. The main steps are template design, answer search, and answer mapping.

The AutoPrompt achieves better performance [51]. It can be summarized as follows: given the original input, a number of additional discrete characters are defined to form a template and the probability of the corresponding [*answers*] is predicted by a pre-trained language model. Continuous prompt converts templates into continuous vectors that can be optimized [52, 53]. Converting templates into vectors that exist in the semantic space facilitates optimization search. The expression of a continuous template is shown in Eq. 2.

$$\mathcal{T} = [x]\left[v_1\right]\left[v_2\right] \ldots \left[v_m\right][mask] \tag{2}$$

where $[v_m]$ is vector.

Prompt-based DA has been widely used. We introduce some representative work here. Chen et al. [22] proposed a framework named GOTTA. They integrated the cloze task. They imitated the main QA task format and efficiently utilized generative-based prompt learning, enabling the model to learn all tasks simultaneously. Liu et al. [23] used prompt learning to extract knowledge from pre-trained models and designed a label-conditioned approach to generate more data with the same labels. In addition, a

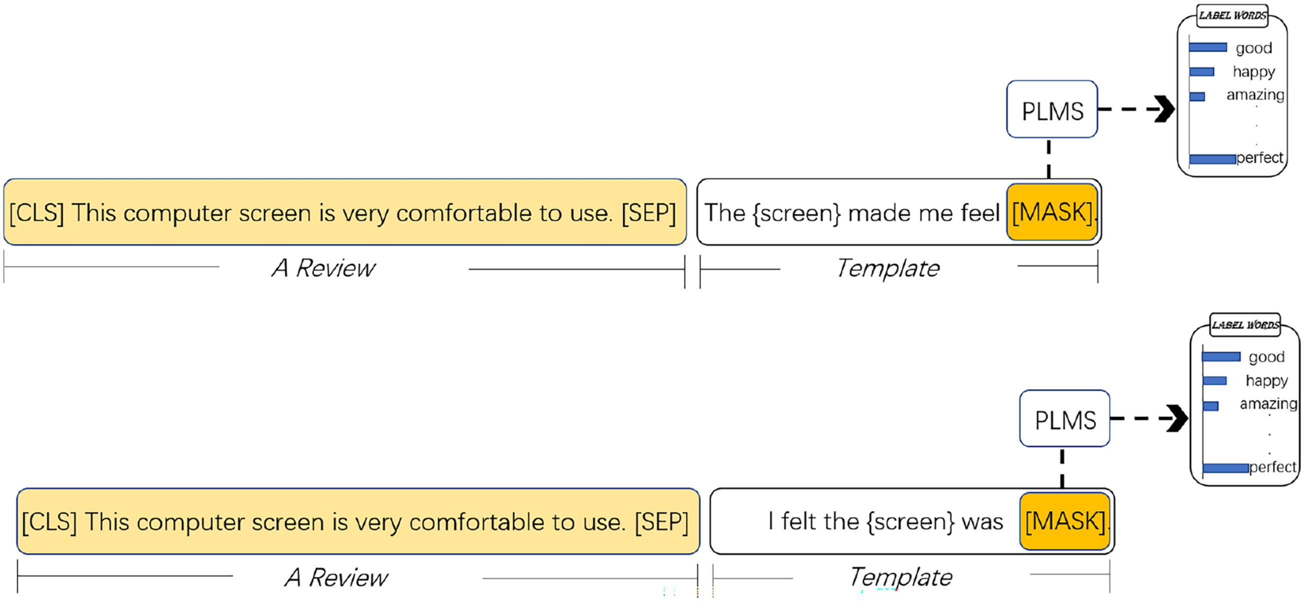
LABEL WORDS
good
happy
amazing
perfect
PLMS
[CLS] This computer screen is very comfortable to use. [SEP]
The {screen} made me feel
[MASK]
A Review
Template
LABEL WORDS
good
happy
amazing
perfect
PLMS
[CLS] This computer screen is very comfortable to use. [SEP]
I felt the {screen} was
[MASK]
A Review
Template

Fig. 3 Schematic diagram of prompt learning

prompt-based QA method was designed to generate new training data from unlabeled text. Wang et al. [24] used a generative-based DA method, utilizing soft prompt and NLU models to filter the generated data from multiple perspectives, ensuring the quality of the newly added data. Abaskohi et al. [25] found that when fine-tuning on small datasets, the model did not perform as well as expected. They combined contrastive learning with prompt tuning and proposed LM-CPPF (Comparative Paraphrasing guided Prompt based Fine-tuning of Language Models) for data augmentation using GPT-3 and OPT-175B.

Through summarizing existing methods, we found that most methods strive to make the model output as close to natural language as possible, but the effect is not satisfactory.

3 Our approach

As mentioned before, the basic idea of the strategies in our study is to label data automatically. Based on the DA techniques described in related work, we believe that the real review can provide more information than the artificially augmented data (e.g. Synonym replacement, Back-translation, etc.). In this section, we first introduce the problem formulation and then describe our augmentation method in detail.

3.1 Problem formulation

Given an initial dataset D, the review text X can be considered as $X = \left[x_1, x_2, \ldots, x_n\right]$ and a label sequence $L = \left[l_1, l_2, \ldots, l_n\right]$ where l_i are the aspects. What we need to do is design an algorithm to identify l_i and give the label (Pos, Neu, Neg) to the X. On this basis, a new dataset with a reasonable proportion is composed of real data and prompt templates filled by pre-trained models.

The first objective of our augmentation task is to assign corresponding aspects and sentiment labels (Pos, Neg) to unlabeled data so that they can be used as training data for the ABSA model.

The second objective of our augmentation task is to use neutral data from Semeval datasets (labeled data) as input, the prompt templates will be filled with the pre-trained model. The filled prompt templates constitute neutral data for the enhanced dataset (label: NEU). As we all know, there are many hard prompt template forms, which ensure the diversity of data. Finally, we get the new dataset which contains the labeled real review (label: Pos, Neg) and filled templates (label: Neu).

3.2 The detail of our model

The framework of our method is shown in Fig. 4. The general procedure for Review2Extreme and Review2Neutral are given in Algorithm 1 and Algorithm 2. The corresponding sub-model framework is shown in the Figs. 5 and 6.

We design two strategies for the problem: Review2Extreme and Review2Neural. They have different responsibilities. The input of Review2Extreme is unlabeled data from YELP, AMAZON, and the preset positive Prompt template. The output is the aspects and corresponding labels (Pos, Neu, Neg). The input of Review2Neural is the Semeval data (the widely-used labeled datasets in ABSA) and normal prompt templates. The output is the templates filled by per-trained models. As shown in Fig. 5. We transformed the task into a probe of unlabeled data by using the prompt template. For example, there is an unlabeled comment: “The food in this restaurant is very bad” This comment is obviously contrary to our preset positive template: “The food in this restaurant made me feel happy”. Finally, we successfully got the data label.

We can decompose the task into the following steps: aspect extraction, automated labeling, and distribution of data in proportion to obtain an enhanced dataset with a balanced distribution. In this paper, we keep the data of corresponding domains in YELP and Amazon, there are a total of five levels from one to five stars in the dataset. We have deleted some data and only selected the reviews x with one and five stars (1★, 5★ in pseudocode) among them, this step is to eliminate the noise of chapter-level labeling. This is a common practice in previous work. Unlike the enhancement strategy for sentence-level sentiment analysis, the DA method for ABSA should ensure that no substitution, splitting, or replacement of aspect words occurs during the process. If the aspects are replaced indiscriminately, the presented noise will interfere locally with the aspect words and affect the sentiment classification effect of the aspect words in the model [54]. So we record the aspects that appear in the public ABSA dataset and select the comments containing the same aspect from the corresponding dataset (YELP, AMAZON). Finally, we obtained the raw data and aspects.

But there is a problem we must solve. For instance, after the first step, we get two sentences including "look": ’It has a bad look’ ’It is so awful, I look for its spare parts but unsuccessfully’. In the first one, "look" is a noun, it means appearance. But in the second one, "look" is a verb that does not have the corresponding sentiment. We must annotate the lexical category to eliminate the noise. This is a sequence labeling problem. As mentioned before, the review text X can be considered as a set of n sentences $X = \left[x_1, x_2, \ldots, x_n\right]$, we define the lexical as $Y = \left[y_1, y_2, \ldots, y_n\right]$. The X and Y can be considered as a Hidden Markov Model (HMM) chain. The HMM model describes the process by which a sentence is produced. X are explicit and Y are implicit.

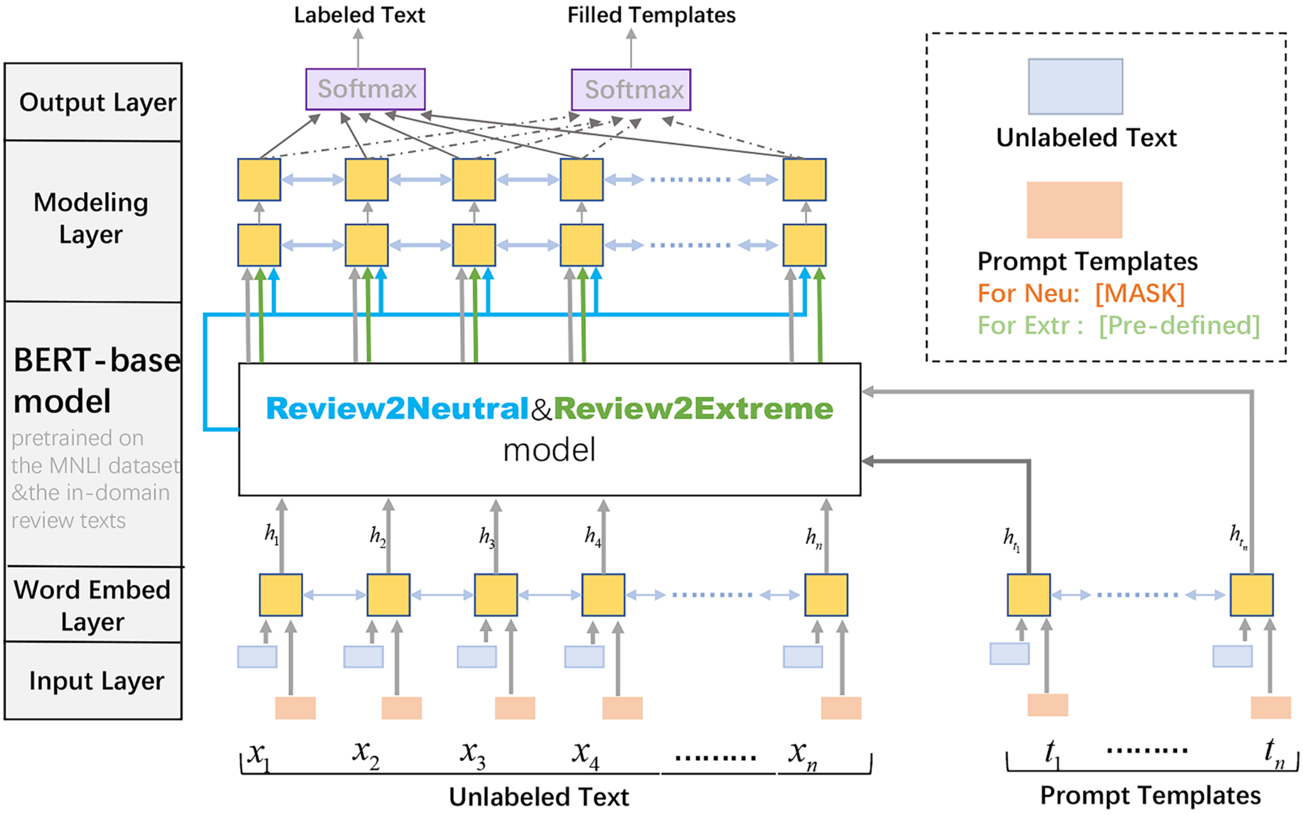

Fig. 4 Framework of our augmentation method

Algorithm 1 Algorithm 1: Review2Extreme (5★, 1★ are rating level in Document-Level dataset D)

```
Input: Document-Level dataset $D = sentence^{5\star,1\star}$;
Output: Labeled dataset $\hat{D} = sentence, aspect, label$;
$\hat{D} = null$;
for $sentence$ in $D$ do
    Aspects ← The aspects in Semeval dataset;
    while $Aspects\ in\ sentence\ is\ Noun$ do
        Labels ← compute the relationship between sentence and positive templates;
    the Labels are voted by multiple templates;
    $\hat{D} \leftarrow \mathcal{D} \cup \{$Aspects , Labels(Pos and Neg)$\}$;
return $\hat{D}$;
```

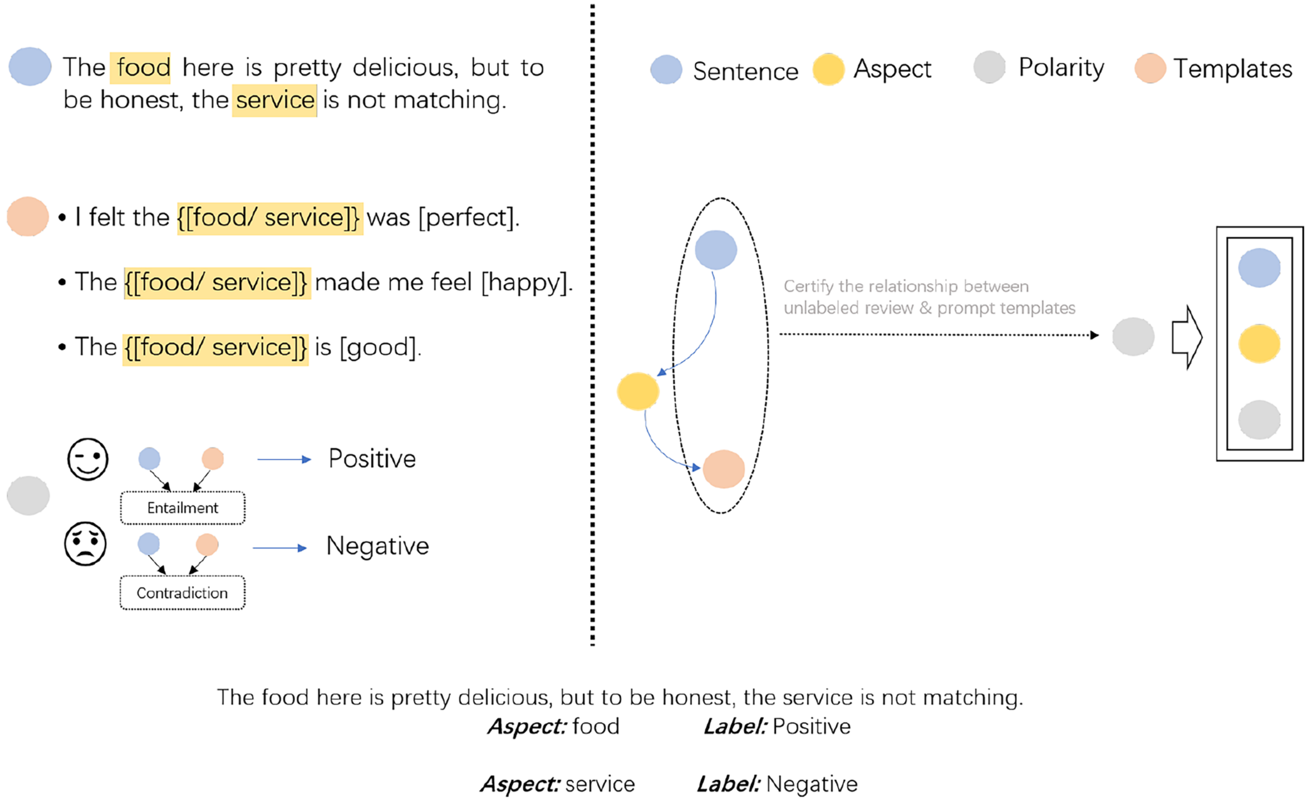
The food here is pretty delicious, but to be honest, the service is not matching.
• I felt the {[food/ service]} was [perfect].
• The {[food/ service]} made me feel [happy].
• The {[food/ service]} is [good].
Entailment
Positive
Contradiction
Negative
Sentence
Aspect
Polarity
Templates
Certify the relationship between unlabeled review & prompt templates
The food here is pretty delicious, but to be honest, the service is not matching.
Aspect: food
Label: Positive
Aspect: service
Label: Negative

Fig. 5 Automated data labeling with Review2Extreme

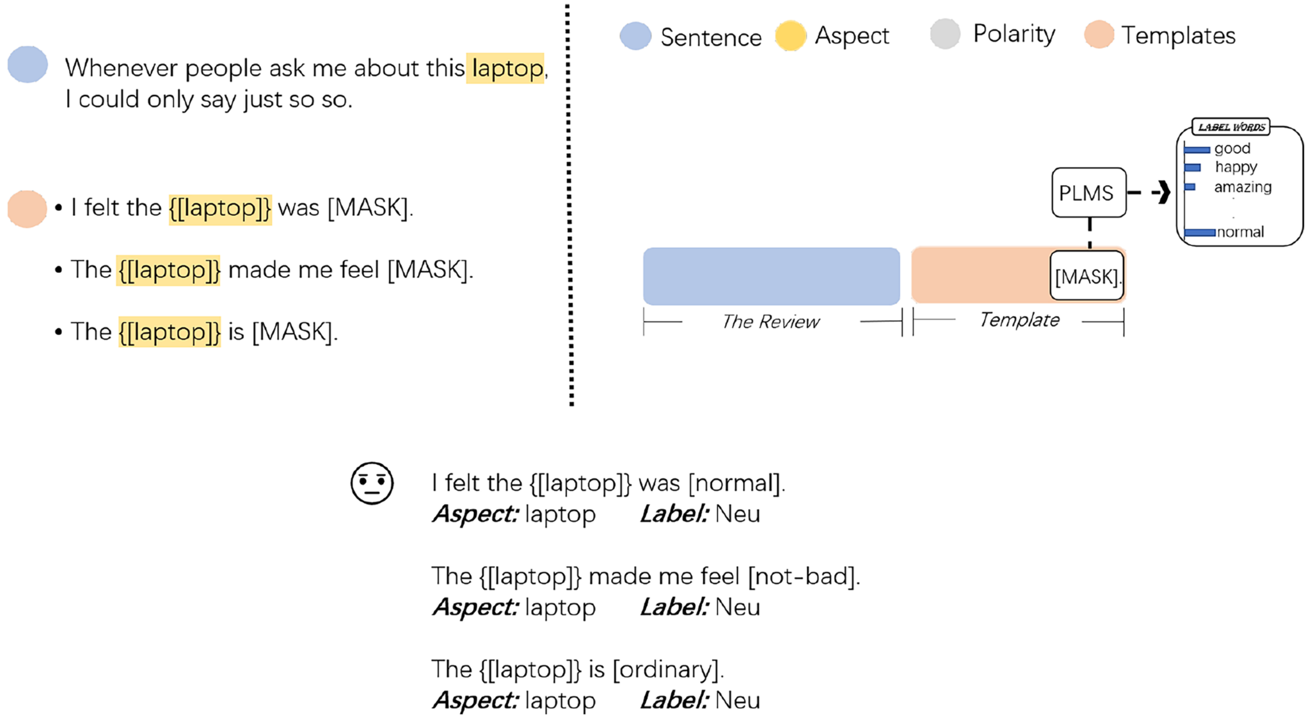
Whenever people ask me about this laptop, I could only say just so so.
• I felt the {[laptop]} was [MASK].
• The {[laptop]} made me feel [MASK].
• The {[laptop]} is [MASK].
Sentence
Aspect
Polarity
Templates
PLMS
LABEL WORDS
good
happy
amazing
normal
[MASK]
The Review
Template
I felt the {[laptop]} was [normal].
Aspect: laptop
Label: Neu
The {[laptop]} made me feel [not-bad].
Aspect: laptop
Label: Neu
The {[laptop]} is [ordinary].
Aspect: laptop
Label: Neu

Fig. 6 Automated data labeling with Review2Neural

Algorithm 2 Algorithm 2: Review2Neutral (sentence$^{\text{neutral}}$ is the Neu data in aspect-level dataset)

```
Input: Semeval dataset $D = sentence^{neutral}$ and corrsponding aspects;
Output: $\hat{\mathbf{D}}^{neutral}$
$\hat{\mathbf{D}}^{neutral} = null$
for $sentence^{neutral}$ in $D$ do
    Labels=Neu;
    Aspects ← the aspects in Semeval datasets;
    Filled templates← The templates filled by PLM according to $D$;
    $\hat{\mathbf{D}}^{neutral}$ ← Filled templates, Aspects , Labels;
return $\hat{\mathbf{D}}^{neutral}$;
```

The implicit states can be transformed following the transition probability. Between X and Y exists the emission probability. For instance, in the sentence "John saw the saw." Odds of producing the sentence "John saw the saw" from the lexical sequence corresponding to the emission probability. We get the equation as follows:

$$\mathrm{P}(\mathrm{x}, \mathrm{y}) = \mathrm{P}(\mathrm{y})\mathrm{P}(\mathrm{x} \mid \mathrm{y}) \tag{3}$$

where

$$\begin{aligned} \mathrm{P}(\mathrm{y}) = \mathrm{P}(\mathrm{PN} \mid \text{ start }) & \\ \times \mathrm{P}(\mathrm{V} \mid \mathrm{PN}) & \\ \times \mathrm{P}(\mathrm{D} \mid \mathrm{V}) & \\ \times \mathrm{P}(\mathrm{N} \mid \mathrm{D}) & \\ \times \mathrm{P}(\text{ end } \mid \mathrm{N}) & \end{aligned} \tag{4}$$

$$\begin{aligned} \mathrm{P}(\mathrm{x} \mid \mathrm{y}) = \mathrm{P}(\text{ John } \mid \mathrm{PN}) & \\ \times \mathrm{P}(\text{ saw } \mid \mathrm{V}) & \\ \times \mathrm{P}(\text{ the } \mid \mathrm{D}) & \\ \times \mathrm{P}(\text{ saw } \mid \mathrm{N}) & \end{aligned}$$

PN means a person's name. *V* means a verb. *D* means an adverb. *N* means a noun. Then we get the equations as follows:

$$\begin{aligned} \mathrm{P}(\mathrm{y}) &= \mathrm{P}\left(\mathrm{y}_1 \mid \text{ start }\right) \times \prod_{\mathrm{l}=1}^{\mathrm{L}-1} \mathrm{P}\left(\mathrm{y}_{\mathrm{l}+1} \mid \mathrm{y}_{\mathrm{l}}\right) \times \mathrm{P}\left(\text{ end } \mid \mathrm{y}_{\mathrm{L}}\right) \\ \mathrm{P}(\mathrm{x} \mid \mathrm{y}) &= \prod_{\mathrm{l}=1}^{\mathrm{L}} \mathrm{P}\left(\mathrm{x}_{\mathrm{l}} \mid \mathrm{y}_{\mathrm{l}}\right) \end{aligned} \tag{5}$$

In Eq. 5, P(y) is defined as transition probability. And P(x) is defined as emission probability.

$\mathrm{P}(y_1 \mid \text{start})$ indicates the odds of the first lexeme chosen at the beginning. $\prod_{\mathrm{l}=1}^{\mathrm{L}-1} \mathrm{P}\left(\mathrm{y}_{\mathrm{l}+1} \mid \mathrm{y}_{\mathrm{l}}\right)$ denotes the probability of a transition probability in the X; $\mathrm{P}(\text{end} \| \ \mathrm{y}_{\mathrm{L}})$ indicates the odds of the last lexeme being at the end. We substitute Eq. 5 into Eq. 3, we get the equations as Eq. 6:

$$\mathrm{P}(\mathrm{x}, \mathrm{y}) = \mathrm{P}\left(\mathrm{y}_1 \mid \text{ start }\right) \prod_{\mathrm{l}=1}^{\mathrm{L}-1} \mathrm{P}\left(\mathrm{y}_{\mathrm{l}+1} \mid \mathrm{y}_{\mathrm{l}}\right) \mathrm{P}\left(\text{ end } \mid \mathrm{y}_{\mathrm{L}}\right) \prod_{\mathrm{l}=1}^{\mathrm{L}} \mathrm{P}\left(\mathrm{x}_{\mathrm{l}} \mid \mathrm{y}_{\mathrm{l}}\right) \tag{6}$$

The emission probability and transition probability are from training data. Substitute it into Eq. 6. We get Eq. 7.

$$\mathrm{P}\left(\mathrm{y}_{l+1} = \mathrm{s}' \mid \mathrm{y}_{\mathrm{l}} = \mathrm{s}\right) = \frac{\operatorname{count}\left(\mathrm{s} \rightarrow \mathrm{s}'\right)}{\operatorname{count}(\mathrm{s})} \tag{7}$$

where the value of $\mathrm{P}(\mathrm{y}_{l+1})$ is the times that y_l appears before y_{l+1}. The same reasoning can be used to prove that $\mathrm{P}\left(\mathrm{x}_{\mathrm{l}} = \mathrm{t} \mid \mathrm{y}_{\mathrm{l}} = \mathrm{s}\right) = \frac{\operatorname{count}(\mathrm{s} \rightarrow \mathrm{t})}{\operatorname{count}(\mathrm{s})}$. The sequence labeling problem can be summarized as Eq. 8.

$$\begin{aligned} \mathrm{y} &= \arg \max_{\mathrm{y} \in \mathrm{Y}} \mathrm{P}(\mathrm{y} \mid \mathrm{x}) \\ &= \arg \max_{\mathrm{y} \in \mathrm{Y}} \frac{\mathrm{P}(\mathrm{x}, \mathrm{y})}{\mathrm{P}(\mathrm{x})} \\ &= \arg \max_{\mathrm{y} \in \mathbb{Y}} \mathrm{P}(\mathrm{x}, \mathrm{y}) \end{aligned} \tag{8}$$

Predicted values are $\tilde{\mathrm{y}} = \arg \max_{\mathrm{y} \in \mathbb{Y}} \mathrm{P}(\mathrm{x}, \mathrm{y})$. We can solve Eq. 8 by the Viterbi algorithm. In the Viterbi algorithm, we introduced two variables δ and ψ, and the maximum probability value for all individual paths $\left(I_1, I_2, \ldots, I_T\right)$ in state i at moment t is defined as Eq. 9.

$$\begin{aligned} &\delta_{t+1}(i) = \max_{1 \leq j \leq N} \left[\delta_t(j) a_{ji}\right] b_i\left(o_{t+1}\right), \\ &\quad i = 1, 2, \ldots, N; t = 1, 2, \ldots, T \end{aligned} \tag{9}$$

We define the $i-1$ node of the path $\left(I_1, I_2, \ldots, I_T\right)$ with the highest probability among all individual paths with state i at moment t as Eq. 10.

Table 1 Statistics of the datasets

	14*laptop*		14*rest*		15*rest*		16*rest*	
	Train	Test	Train	Test	Train	Test	Train	Test
Pos	970	341	2151	725	911	331	1234	468
Neu	455	169	632	196	36	36	70	31
Neg	843	127	794	196	261	192	457	123
Ratio	2.1:1:1.9	2.7:1.3:1	3.4:1:1.2	3.7:1:1	25.3:1:7.3	9.2:1:5.3	17.8:1:6.5	15.1:1:4

$$\psi_t(i) = \arg\max_{1\leq j\leq N}\left[\delta_{t-1}(j)a_{ji}\right], i = 1, 2, \dots, N; t = 1, 2, \dots, T \tag{10}$$

The algorithm initialization can be shown as Eq. 11:

$$\begin{aligned} &\delta_1(i) = \pi_i b_i\left(o_1\right), \quad i = 1, 2, \ \mathrm{L}, N \\ &\psi_1(i) = 0, \quad i = 1, 2, \ \mathrm{L}, N \end{aligned} \tag{11}$$

As shown in Eq. 12, for $t = 2, \dots, T$, we have the equation:

$$\begin{aligned} &\delta_t(i) = \max_{1\leq j<N}\left[\delta_{t-1}(j)a_{ji}\right]b_i\left(o_t\right), i = 1, 2, \ \mathrm{L}, N \\ &\psi(i) = \arg\max_{1\leq j<N}\left[\delta_{t-1}(j)a_{jt}\right], \quad i = 1, 2, \ \mathrm{L}, N \end{aligned} \tag{12}$$

The algorithm will be ended when $P^* = \max_{1\leq i\leq N} \delta_T(i)$ and $i_T^* = \arg\max_{1\leq i\leq N}\left[\delta_T(i)\right]$. We can find the optimal path by performing optimal path backtracking, for $t = T-1, T-2, \dots, 1$, it is shown as Eq. 13:

$$\begin{aligned} &i_t^* = \psi_{t+1}\left(i_{t+1}^*\right) \\ &I^* = \left(i_1^*, i_2^*, \ \mathrm{L}, i_T^*\right) \end{aligned} \tag{13}$$

We further leverage the self-consistency mechanism [2, 55] to consolidate the labeling correctness. Specifically, for each of the three prompt templates, we set the PLM decoder to generate answers independently, and we select the one with high voting consistency.

4 Experiments

4.1 Datasets

SemEval datasets [56, 57] have been widely used in ABSA for many years. It can be seen from the dataset that the amount of data has been difficult to fit the requirements of large models. The details are shown in Table 1.

The Amazon review dataset records user reviews of products on Amazon.com and is a classic dataset for recommendation systems, and Amazon is always updating this dataset.

The YELP dataset includes 4.7 million user reviews and 12 metropolitan areas. It also covers 1 million tips from 1.1 million users, and over 1.2 million merchant attributes (such as hours of operation, availability of parking, reservation availability, and environment information). The data in YELP and Amazon is chapter-level data, they are divided into five sentiment tendencies from one to five stars.

4.2 Experimental models

ASGCN [58]: ASGCN is based on GCN and contextual information about the word order is captured starting from the LSTM layer. A multi-layer graph convolution structure is then implemented on top of the LSTM output to obtain aspect features.

CABASC [59]: CABASC is based on a sentence-based content attention mechanism that embeds sentences and aspects separately. This allows extracting the important information of specific aspect words in a sentence from a global perspective, taking into account the location and relevance of the information. A contextual attention mechanism is designed in CABASC to generate custom memory blocks of aspect words in each sentence considering the order and relevance between words and aspect words.

LSTM [60]: LSTM is a special type of recurrent neural network (RNN) for modeling time-series data. Two fully concatenated layers and a softmax layer form the main network structure. The ASGCN, CABASC, and LSTM are also used in Li et al. (DRAWS and PWSS) [54].

R-GAT+BERT [61]: R-GAT solves this problem by effectively encoding grammatical information. Firstly, by refactoring and pruning the ordinary dependency parsing tree, a unified aspect-oriented approach was defined based on the aspect The dependency tree structure of. Based on the graph attention mechanism, encode a new tree structure for data label prediction.

Table 2 R-GAT performance comparison among none and ours

Model		14_*laptop*		14_*rest*	
		None	Ours	None	Ours
R-GAT	Acc	83.26	**85.37**	77.16	**77.82**
	Macro-F1	76.03	**77.58**	73.81	**74.26**
R-GAT+BERT	Acc	86.48	**87.22**	78.01	**79.18**
	Macro-F1	81.37	**81.96**	74.25	**75.14**

Table 3 Comparison of the performance Macro-F1 (%) among None, PWSS [54], DRAWS [54] and Ours for different training sizes

Model		ASGCN				CABASC				LSTM			
Dataset		None	PWSS	DRAWS	Ours	None	PWSS	DRAWS	Ours	None	PWSS	DRAWS	Ours
14_*laptop*	50	36.36	41.37	36.25	**37.88**	39.79	45.03	51.39	**57.04**	36.09	41.08	47.98	**53.61**
	500	51.59	54.12	57.11	**59.79**	54.34	53.99	60.61	**62.58**	50.67	55.26	59.75	**64.30**
	full	59.17	59.27	61.17	**62.39**	57.26	61.2	63.69	**65.21**	50.54	57.82	62.05	**68.47**
14_*rest*	50	33.13	40.44	37.53	**41.43**	27.59	43.44	47.58	**49.52**	26.26	41.62	37.78	**46.74**
	500	51.21	55.23	57.94	**61.52**	51.09	55.61	59.92	**64.37**	55.17	54.06	62.53	**64.41**
	full	55.81	60.62	64.99	**67.21**	59.89	60.95	64.83	**68.17**	57.89	61.58	63.88	**66.03**
15_*rest*	50	24.57	42.23	31.56	**39.70**	24.80	39.91	43.40	**44.63**	24.77	40.72	28.80	**35.46**
	500	56.80	51.57	51.26	**54.37**	45.70	51.62	53.03	**53.44**	45.47	50.30	54.18	**54.69**
	full	47.60	55.39	52.64	**57.86**	49.48	52.57	54.26	**55.83**	47.28	52.48	53.23	**56.21**
16_*rest*	50	31.55	44.24	37.72	**44.01**	30.38	40.86	37.86	**40.44**	28.42	40.67	32.51	**37.95**
	500	52.34	54.01	54.91	**56.32**	49.28	52.04	54.78	**57.53**	49.61	50.54	53.08	**54.32**
	full	53.12	57.87	57.35	**58.64**	48.68	57.86	55.71	**63.57**	49.62	56.02	55.56	**59.73**

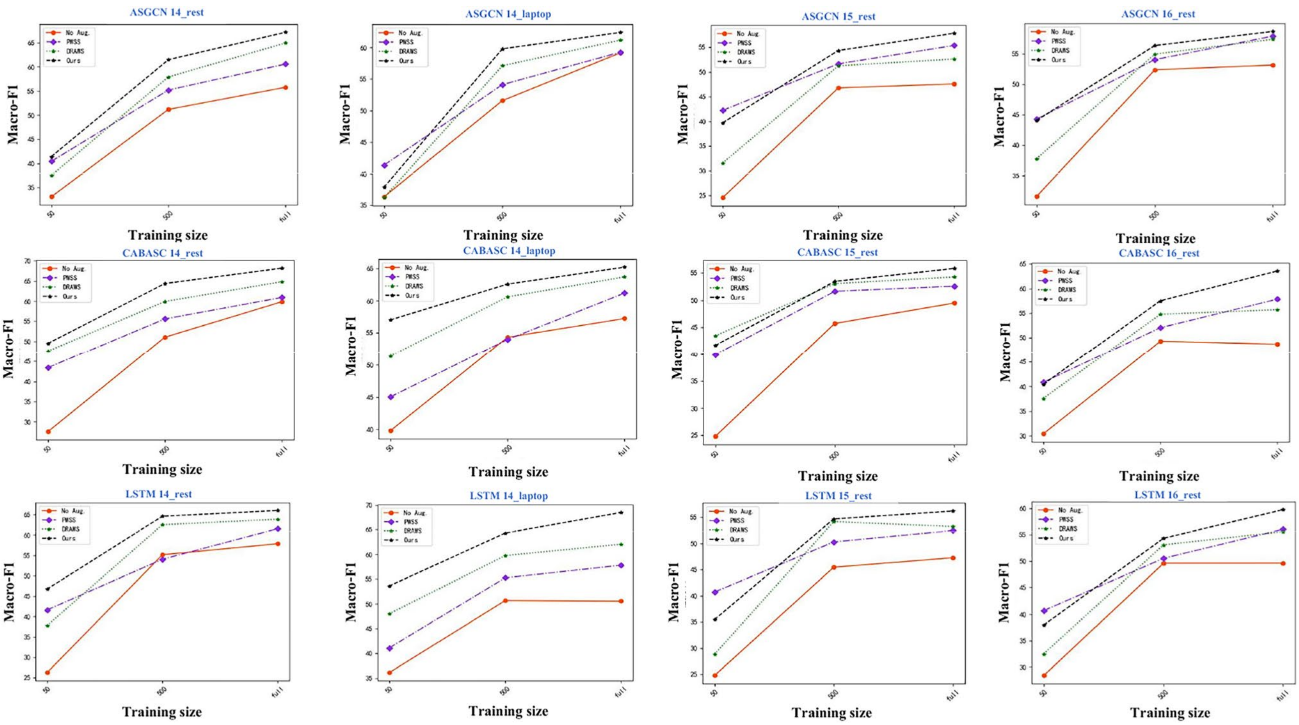

Fig. 7 Comparison of macro-F1(%) on different training sizes among none, PWSS [54], DRAWS [54] and ours

4.3 Baseline models

DRAWS and PWSS [54]: Pos-Wise Synonym Substitution (PWSS) and Dependency Relation-based Word Swap (DRAWS) are proposed by Guangmin Li et al. PWSS selects synonyms from general dictionaries for substitution. PWSS enables a reasonable increase in the capacity of the training set. DRAWS have a better ability of sentence semantics on polarity orientation (positive, negative, and neutral).

4.4 Setting of experiment

To make maximum use of the knowledge in the pre-trained model, we selected three templates for the templates by adding them to the text. The number of templates can be changed arbitrarily.

- I felt the [aspect] was [MASK].
- The [aspect] made me feel [MASK].

- The [aspect] is [MASK].

where [aspect] is the placeholder for the aspect term, and [MASK] represents the masked word for BERT which is pre-trained on the MNLI dataset.

We use the weights released by Morris et al. [62], which were trained on the MNLI dataset. The models are trained for 20 epochs. According to [63], we finetune the Prompt-based DA method until the training losses are around 1e-07 to get a stable model.

As Li et al. [54] described, according to the original proportions of polarity in the training set, 50, 500, and all instances are selected in turn while the data capacity of the test set and validation set remains unchanged.

4.5 Results and discussion

The additional experimental results based are shown in Table 2. The performance for each model is shown in Table 3 and illustrated in Figure 7. Our experimental setup was consistent with the baseline models. Specifically, the experimental models ASGCN, CABASC, and LSTM were trained for different training sizes in the baseline model [54].

Our method achieves better results than the baseline model on all experimental models. The following observations are made. In terms of the training set size, the classification performance of the three models shows an upward trend with the increase in data capacity. The experimental results confirm that the augmentation strategies are effective. Especially in the training results of the training size from 50 to 500, the Macro-F1 is increased by a large margin. It is also seen that the classification results are slowly improved from training size 500 to full.

This result ties well with the previous study [64, 65]. This is attributed to the injection of noise information with the increase of the data scale. The results of this experiment show that the model gain is not simply linearly related to the amount of data. At the same time, it also illustrates the necessity of data correction and textual data noise removal for existing datasets. This needs to be verified by further experiments. Unlike existing methods, our approach provides real data. Although the Semeval dataset is also composed of real data, our data has less noise compared to manual annotation, resulting in better performance than the original data. Furthermore, it can be seen from Table 1 that the class distributions are dramatically imbalanced on datasets 15_rest and 16_rest. The data we annotate can be added to relevant categories to balance the distribution of the data. Further improved the performance of the ABSA models. The baseline DRAWS, base1 and PWSS [54] did not verify their methods on pre-trained models. In order to better demonstrate our method, we supplemented the relevant experiments of the pre-rained models. We conducted experiments based on the code published by the authors. The results are shown in Table 2.

In Table 4, we present the case study to show the model effectively. R1 is difficult for our models. As we can see, there are confusing words in R1 (food was exceeding expectations), but there is a prerequisite (when they actually gave people the meals they ordered), so honestly speaking, it is a negative review. This case demonstrates that our model has some shortcomings. In a few situations, the model's overall grasp of the information in the corpus is inadequate. Hand-constructed templates do not efficiently apply the knowledge present in pre-trained models. R2 indicates that our method can effectively label data in most cases. R3 and R4 are implicit sentiment reviews. These cases demonstrate that our model has good recognition of implicitly expressed sentiment. R5 and R6 indicate that different prepositions can affect the performance of the model. Prepositions that represent semantic progression (e.g. 'also') may mislead the model into ignoring subsequent information, thereby affecting the performance of the model. R7 is a template that is filled by per-trained models. We select many templates for NEU data, the aspect is the same as Semeval datasets. Our method can perform data completion on datasets with imbalanced label distribution, as well as bias correction on existing datasets. As a low computational resource method, it has

Table 4 Case study

Review	Truth	Ours
1: The absolute worst service I've ever experienced and the **food** was exceeding expectations. (when they actually gave people the meals they ordered)	Neg	Pos
2: The absolute worst **service** I've ever experienced and the food was exceeding expectations. (when they actually gave people the meals they ordered)	Neg	Neg
3: The **sushi** is cut in blocks bigger than my cell phone	Neg	Neg
4: **Lunch** came with pickles and slaw, no extra charge	Pos	Pos
5: The **dinner** here prices are a bit over the top. Also taste good	Pos	Neg
6: At first, I felt that the **computer** at this price would not perform well, but its performance greatly exceeded my expectations	Pos	Pos
7: The **service** makes me feel [OK], although there is still room for improvement, it is already not bad	Neu	Neu

Table 5 Comparison of the macro-F1(%) (no Review2Extreme : *Sem_Ext* + *New_Neu*)

Dataset	ASGCN		CABASC		LSTM	
14_*laptop*	62.39	**61.87**	65.21	**64.87**	68.47	**67.24**
14_*rest*	67.21	**66.31**	68.17	**67.74**	66.03	**65.17**
15_*rest*	57.86	**56.64**	55.83	**54.37**	56.21	**55.18**
16_*rest*	58.64	**57.39**	63.57	**61.93**	59.73	**57.67**

Table 6 Comparison of the Macro-F1(%) (No Review2Neutral : *New_Ext* + *Sem_Neu*)

Dataset	ASGCN		CABASC		LSTM	
14_*laptop*	62.39	**60.22**	65.21	**58.82**	68.47	**63.14**
14_*rest*	67.21	**64.60**	68.17	**65.47**	66.03	**63.26**
15_*rest*	57.86	**53.19**	55.83	**51.54**	56.21	**52.36**
16_*rest*	58.64	**53.41**	63.57	**58.02**	59.73	**55.57**

great application prospects. It is undeniable that there is still a lot of room for improvement in our method, such as the complexity of human language (such as prepositions with diverse expressions) and the erroneous judgments caused by multiple aspects coexisting in a sentence, which urgently need to be addressed.

Based on the experimental results, we have listed the advantages and disadvantages of our method and other classic data augmentation techniques(including baseline) for a more detailed comparison:

The overview of Wei et al. [8]: Four data augmentation techniques have been proposed, including synonym replacement, random insertion, random exchange, and random deletion. A comparative study was conducted on text classification experiments using five datasets on deep learning models RNN and CNN.

The advantages of Wei et al. [8]: For the first time, random inserts, swaps, and deletions were used for data augmentation, and their methods were validated on convolutional neural networks and recurrent neural networks; The method is simple, intuitive and easy to understand; Under appropriate parameters, each raw data can generate nine enhanced sentences, and in most cases, the original data labels are retained.

The disadvantages of Wei et al. [8]: Firstly, the output of the EDA model has the possibility of changing semantics, which may conflict with the original labels and provide incorrect learning samples for the ABSA model. Secondly, pre-trained models have become the mainstream of current NLP research, and simple DA is no longer able to substantially improve model performance [36]. Finally, according to the research of Ebrahimi et al. [15]., small changes in training data have a significant impact on the performance of the ABSA model. The uncontrollability of EDA (such as no improvement in temporary analysis when replacing heads with synonyms [66]) may bring unknown data noise.

Table 7 Comparison of the performance(%) (no Review2Extreme : *Sem_Ext* + *New_Neu*)

Model		14_*laptop*		14_*rest*	
		Overall	Ablation	Overall	Ablation
R-GAT	Acc	85.37	**84.63**	77.82	**77.20**
	Macro-F1	77.58	**76.28**	74.26	**74.01**
R-GAT+BERT	Acc	87.22	**87.09**	79.18	**78.57**
	Macro-F1	81.96	**81.62**	75.14	**74.69**

The overview of Li et al. [54]: It is a method based on EDA that utilizes part of speech, external domain knowledge, and syntactic dependencies to achieve DA through synonym replacement and dependency-based word exchange. These strategies were evaluated through extensive experiments using three representative deep learning models—ASGCN, CABASC, and LSTM—on four common datasets (Semeval dataset).

The advantages of Li et al. [54]: Unlike previous sentiment-based analysis methods, this method combines DA with dependency parsing trees, incorporating external knowledge to improve data quality. An analysis was conducted on common deep-learning network architectures. The method proposed by the author enhances the generalization ability of the model.

The disadvantages of Li et al. [54]: Adding noise information that interferes with real labels during word replacement cannot balance semantic integrity and syntactic correctness. The data constructed by the method is still an extension based on old data, which can only bring limited new information to the model.

The overview of ours: Automated annotation of real unlabeled data based on prompt learning. By constructing a rich hard prompt template, based on the labeled data of NEU, the filled template is introduced back into the model. Due to the fact that the existing NEU data is manually annotated,

Table 8 Comparison of the performance(%) (No Review2Neutral : *New_Ext* + *Sem_Neu*)

Model		14_*laptop*		14_*rest*	
		Overall	Ablation	Overall	Ablation
R-GAT	Acc	85.37	**83.28**	77.82	**76.42**
	Macro-F1	77.58	**76.19**	74.26	**73.90**
R-GAT+BERT	Acc	87.22	**86.53**	79.18	**78.13**
	Macro-F1	81.96	**81.44**	75.14	**74.27**

ablation experiments have shown that it contains a significant amount of data noise. This operation further corrects the label bias of the existing NEU data. This method can also quantitatively introduce data into imbalanced datasets, improving the effectiveness of the ABSA model.

The advantages of ours: The method occupies less computational resources, is intuitive and easy to understand, and is simple to use; Provides authentic data while balancing semantic integrity and syntactic correctness; Provides controllable data augmentation methods.

The disadvantages of ours: The discriminative ability for complex statements needs to be strengthened; Compared to the soft prompt template, the hard prompt template cannot fully utilize the knowledge in the pre-trained model; We should enhance research on semantic and grammatical structural information to further improve model performance. Eliminate data noise.

4.6 Ablation study

4.6.1 Review2Extreme and Review2Neutral's respective influence

As mentioned before, our approach consists of two parts. We use the following symbols to represent the corresponding data:

Sem_Neu: Raw neutral data in the Semeval dataset

Sem_Ext: Raw positive and negative data in the Semeval dataset

New_Neu: The new neutral data

New_Ext: The new positive and negative data we labeled

In this subsection, we attempt to block Review2Extreme and Review2Neutral separately to observe their influence on the performance of different ABSA models. The experimental results are shown in Tables 5, 6, 7, 8. Through the analysis of the experimental results, we believe that there are

Table 9 Comparison of the Macro-F1(%) (determined aspects/unified aspect)

Model (determined/ unified aspect)	ASGCN		CABASC		LSTM	
14_*laptop*	60.22	**57.35**	58.82	**55.49**	63.14	**59.08**
14_*rest*	64.60	**54.62**	65.47	**54.47**	63.26	**54.76**
15_*rest*	53.19	**50.66**	51.54	**50.27**	52.36	**49.36**
16_*rest*	53.41	**50.39**	58.02	**55.11**	55.57	**52.34**

Table 10 Comparison of the R-GAT (%) (determined aspects/unified aspect)

Model		14_*laptop*		14_*rest*	
		Determined	Unified	Determined	Unified
R-GAT	Acc	83.28	**77.63**	76.42	**69.90**
	Macro-F1	76.19	**70.25**	73.90	**67.88**
R-GAT+BERT	Acc	86.53	**79.87**	78.13	**71.24**
	Macro-F1	81.44	**73.92**	74.27	**69.25**

Table 11 Comparison of the macro-F1(%) (multiple templates/random single prompt)

Model (multiple/single prompt)	ASGCN		CABASC		LSTM	
14_*laptop*	60.22	**55.98**	58.82	**57.11**	63.14	**58.96**
14_*rest*	64.60	**57.19**	65.47	**58.87**	63.26	**53.12**
15_*rest*	53.19	**52.02**	51.54	**49.23**	52.36	**50.59**
16_*rest*	53.41	**51.81**	58.02	**55.42**	55.57	**52.84**

Table 12 Comparison of the R-GAT (%) (multiple templates/random single prompt)

Model		14_*laptop*		14_*rest*	
		Multiple	Single	Multiple	Single
R-GAT	Acc	83.28	**76.40**	76.42	**71.13**
	Macro-F1	76.19	**69.22**	73.90	**68.35**
R-GAT+BERT	Acc	86.53	**81.07**	78.13	**72.09**
	Macro-F1	81.44	**75.93**	74.27	**69.87**

some mistakes in Neu data which is from the Semeval dataset, it has a certain negative influence on the performance of the ABSA model.

4.6.2 Aspect independent prompts

As mentioned before, our approach computes the relationship between the prompt templates and unlabeled data and then assigns labels to expand the training data. In this subsection, we will replace different aspects with a unified aspect. (e.g. The [aspect] made me feel good. transforms to This made me feel good. The performance is shown in Tables 9 and 10, we found that this operation will cause a significant decrease in the quality of newly labeled data. Indicating that the determined aspects have a crucial impact on whether the model can correctly label data.)

4.6.3 Single prompt template

As mentioned before, the labels (positive and negative) are voted upon by multiple templates. In this subsection, we attempt to assign the corresponding labels by one template. The performance is shown in Tables 11 and 12, the experimental results indicate that multiple templates can effectively correct errors in the model and improve the quality of the dataset.

5 Conclusions and future work

In this paper, we systematically explore the effects of data augmentation for ABSA and propose a novel prompt learning-based data augmentation method, which allows us to compute the relationship between prompt templates and unlabeled data, leverage them to enrich data resources and improve generalization capability via prompt learning. The experimental results on four well-studied datasets demonstrate that our model not only achieves results on par with existing state-of-the-art data augmentation methods on a few occasions but also significantly outperforms existing ABSA base models and data augmentation methods on most occasions, indicating its strong robustness in various base ABSA models, including ASGCN [58], CABASC [59] and etc. Further explorations on incorporating more types of soft template (soft prompt is optimized in vector space) and data augmentation methods for enhancing the performance of ABSA will be addressed in our future work.

Acknowledgements This study was supported by the National Natural Science Foundation of China (62176074).

Data availability The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

References

1. Bu K, Liu Y, Ju X (2024) Efficient utilization of pre-trained models: a review of sentiment analysis via prompt learning. Knowl-Based Syst 283:111148. https://doi.org/10.1016/j.knosys.2023.111148
2. Li Z, Zou Y, Zhang C, Zhang Q, Wei Z (2021) Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training. In: Proceedings of the 2021 conference on empirical methods in natural language processing. Association for computational linguistics, Online and Punta Cana, Dominican Republic, pp 246–256. https://doi.org/10.18653/V1/2021.EMNLP-MAIN.22
3. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Bartlett PL, Pereira FCN, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in neural information processing systems 25: 26th Annual Conference on Neural Information Processing Systems. Proceedings of a Meeting Held 3–6 Dec 2012, Lake Tahoe, Nevada, United States, pp 1106–1114 . https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html
4. Perez L, Wang J (2017) The effectiveness of data augmentation in image classification using deep learning. CoRR. arXiv:1712.04621
5. Singh J, McCann B, Keskar NS, Xiong C, Socher R (2019) XLDA: Cross-lingual data augmentation for natural language inference and question answering. CoRR. arXiv:1905.11471
6. Fadaee M, Bisazza A, Monz C (2017) Data augmentation for low-resource neural machine translation, pp 567–573. https://doi.org/10.18653/V1/P17-2090
7. Mueller J, Thyagarajan A (2016) Siamese recurrent architectures for learning sentence similarity. Proceedings of the AAAI conference on artificial intelligence. 30(1). https://doi.org/10.1609/aaai.v30i1.10350
8. Wei JW, Zou K (2019) EDA: easy data augmentation techniques for boosting performance on text classification tasks, pp 6381–6387. https://doi.org/10.18653/v1/D19-1670
9. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville AC, Bengio Y (2014) Generative adversarial nets. Adv Neural Info Process Syst pp. 27
10. Kingma DP, Welling M (2014) Auto-encoding variational bayes. The International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=33X9fd2-9FyZd
11. Gupta R (2019) Data augmentation for low resource sentiment analysis using generative adversarial networks, pp 7380–7384. https://doi.org/10.1109/ICASSP.2019.8682544

12. Bayer M, Kaufhold M, Reuter C (2023) A survey on data augmentation for text classification. ACM Comput Surv 55(7):146–114639. https://doi.org/10.1145/3544558
13. Kobayashi S (2018) Contextual augmentation: data augmentation by words with paradigmatic relations, pp 452–457. https://doi.org/10.18653/v1/n18-2072
14. Wang B, Ding L, Zhong Q, Li X, Tao D (2022) A contrastive cross-channel data augmentation framework for aspect-based sentiment analysis. In: Proceedings of the 29th International-Conference on Computational Linguistics, Int Committee Comput Linguist, Gyeongju, Republic of Korea, 6691–6704
15. Ebrahimi J, Rao A, Lowd D, Dou D, Hotflip (2018) White-box adversarial examples for text classification, pp 31–36. https://doi.org/10.18653/V1/P18-2006
16. Li B, Hou Y, Che W (2022) Data augmentation approaches in natural language processing: a survey. AI Open 3:71–90. https://doi.org/10.1016/j.aiopen.2022.03.001
17. Johnson JM, Khoshgoftaar TM (2019) Survey on deep learning with class imbalance. J Big Data 6:27. https://doi.org/10.1186/S40537-019-0192-5
18. López V, Fernández A, García S, Palade V, Herrera F (2013) An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. Inf Sci 250:113–141. https://doi.org/10.1016/J.INS.2013.07.007
19. Thabtah FA, Hammoud S, Kamalov F, Gonsalves AH (2020) Data imbalance in classification: experimental evaluation. Inf Sci 513:429–441. https://doi.org/10.1016/J.INS.2019.11.004
20. Bu J, Daw A, Maruf M, Karpatne A (2021) Learning compact representations of neural networks using discriminative masking (DAM), pp 3491–3503
21. Liu Z, Li J, Shen Z, Huang G, Yan S, Zhang C (2017) Learning efficient convolutional networks through network slimming, pp 2755–2763. https://doi.org/10.1109/ICCV.2017.298
22. Chen X, Zhang Y, Deng J, Jiang J, Wang W Gotta (2023) Generative few-shot question answering by prompt-based cloze data augmentation, pp 909–917 . https://doi.org/10.1137/1.9781611977653.CH102
23. Liu J, Chen Y, Xu J (2022) Low-resource NER by data augmentation with prompting, pp 4252–4258. https://doi.org/10.24963/IJCAI.2022/590
24. Wang Y, Xu C, Sun Q, Hu H, Tao C, Geng X, Jiang D (2022) Promda: prompt-based data augmentation for low-resource NLU tasks, pp 4242–4255. https://doi.org/10.18653/V1/2022.ACL-LONG.292
25. Abaskohi A, Rothe S, Yaghoobzadeh Y (2023) LM-CPPF: paraphrasing-guided data augmentation for contrastive prompt-based few-shot fine-tuning, pp 670–681. https://doi.org/10.18653/V1/2023.ACL-SHORT.59
26. Huang X, Li J, Wu J, Chang J, Liu D (2023) Transfer learning with document-level data augmentation for aspect-level sentiment classification. IEEE Trans Big Data 9(6):1643–1657. https://doi.org/10.1109/TBDATA.2023.3310267
27. Zhao H, Huang L, Zhang R, Lu Q, Xue H (2020) Spanmlt: a span-based multi-task learning framework for pair-wise aspect and opinion terms extraction, pp 3239–3248. https://doi.org/10.18653/v1/2020.acl-main.296
28. Chen S, Liu J, Wang Y, Zhang W, Chi Z (2020) Synchronous double-channel recurrent network for aspect-opinion pair extraction, pp 6515–6524. https://doi.org/10.18653/v1/2020.acl-main.582
29. Chen Z, Qian T (2020) Relation-aware collaborative learning for unified aspect-based sentiment analysis, pp 3685–3694. https://doi.org/10.18653/v1/2020.acl-main.340
30. Luo H, Ji L, Li T, Jiang D, Duan N (2020) GRACE: gradient harmonized and cascaded labeling for aspect-based sentiment analysis. EMNLP 2020, pp 54–64. https://doi.org/10.18653/v1/2020.findings-emnlp.6
31. Cai H, Tu Y, Zhou X, Yu J, Xia R (2020) Aspect-category based sentiment analysis with hierarchical graph convolutional network, pp 833–843. https://doi.org/10.18653/v1/2020.coling-main.72
32. Li Y, Yang Z, Yin C, Pan X, Cui L, Huang Q, Wei T (2020) A joint model for aspect-category sentiment analysis with shared sentiment prediction layer 12522:388–400. https://doi.org/10.1007/978-3-030-63031-7_28
33. Liu J, Teng Z, Cui L, Liu H, Zhang Y (2021) Solving aspect category sentiment analysis as a text generation task, pp 4406–4416. https://doi.org/10.18653/v1/2021.emnlp-main.361
34. Li R, Chen H, Feng F, Ma Z, Wang X, Hovy EH (2021) Dual graph convolutional networks for aspect-based sentiment analysis, pp 6319–6329. https://doi.org/10.18653/V1/2021.ACL-LONG.494
35. Zhong Q, Ding L, Liu J, Du B, Jin H, Tao D (2023) Knowledge graph augmented network towards multiview representation learning for aspect-based sentiment analysis. IEEE Trans Knowl Data Eng 35(10):10098–10111. https://doi.org/10.1109/TKDE.2023.3250499
36. Longpre S, Wang Y, DuBois C (2020) How effective is task-agnostic data augmentation for pretrained transformers? EMNLP 2020, pp 4401–4411. https://doi.org/10.18653/V1/2020.FINDINGS-EMNLP.394
37. LeCun Y, Bottou L, Haffner BYP (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324. https://doi.org/10.1109/5.726791
38. Coulombe C (2018) Text data augmentation made simple by leveraging NLP cloud apis. arXiv:1812.04718
39. Belinkov Y, Bisk Y (2018) Synthetic and natural noise both break neural machine translation. Int Conf Learn Represent. https://openreview.net/forum?id=BJ8vJebC-
40. Feng SY, Gangal V, Kang D, Mitamura T, Hovy EH (2020) Genaug: data augmentation for finetuning text generators. arXiv:2010.01794
41. Wang WY, Yang D (2015) That's so annoying!!!: a lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets, pp 2557–2563. https://doi.org/10.18653/V1/D15-1306
42. Marivate V, Sefara T (2020) Improving short text classification through global augmentation methods 12279:385–399. https://doi.org/10.1007/978-3-030-57321-8_21
43. Rizos G, Hemker K, Schuller BW (2019) Augment to prevent: short-text data augmentation in deep learning for hate-speech classification, pp 991–1000. https://doi.org/10.1145/3357384.3358040
44. Huong TH, Hoang VTA (2020) Data augmentation technique based on text for vietnamese sentiment analysis, pp 13–1135. https://doi.org/10.1145/3406601.3406618
45. Wu X, Lv S, Zang L, Han J, Hu S (2019) Conditional BERT contextual augmentation 11539, pp 84–95. https://doi.org/10.1007/978-3-030-22747-0_7
46. Hu Z, Tan B, Salakhutdinov R, Mitchell TM, Xing EP (2019) Learning data manipulation for augmentation and weighting, pp 15738–15749
47. Qu Y, Shen D, Shen Y, Sajeev S, Chen W, Han J (2021) Coda: contrast-enhanced and diversity-promoting data augmentation for natural language understanding. The International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.2010.08670. https://openreview.net/forum?id=Ozk9MrX1hvA
48. Anaby-Tavor A, Carmeli B, Goldbraich E, Kantor A, Kour G, Shlomov S, Tepper N, Zwerdling N (2020) Do not have enough

data? deep learning to the rescue!, pp 7383–7390. https://doi.org/10.1609/AAAI.V34I05.6233

49. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D.M, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D (2020) Language models are few-shot learners
50. Schick T, Schütze H (2021) Exploiting cloze-questions for few-shot text classification and natural language inference, pp 255–269. https://doi.org/10.18653/v1/2021.eacl-main.20
51. Shin T, Razeghi Y, IV RLL, Wallace E, Singh S (2020) Autoprompt: eliciting knowledge from language models with automatically generated prompts, pp 4222–4235. https://doi.org/10.18653/v1/2020.emnlp-main.346
52. Liu X, Zheng Y, Du Z, Ding M, Qian Y, Yang Z, Tang J GPT understands, too. CoRR **abs/2103.10385** (2021) arXiv:2103.10385
53. Gu Y, Han X, Liu Z, Huang M (2022) PPT: pre-trained prompt tuning for few-shot learning, pp 8410–8423. https://doi.org/10.18653/v1/2022.acl-long.576
54. Li G, Wang H, Ding Y, Yan ZKX (2023) Data augmentation for aspect-based sentiment analysis. Int J Mach Learn Cybern 14(1):125–133. https://doi.org/10.1007/S13042-022-01535-5
55. Wang X, Wei J, Schuurmans D, Le QV, Chi EH, Narang S, Chowdhery A, Zhou D (2023) Self-consistency improves chain of thought reasoning in language models. The International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=1PL1NIMMrw
56. Pontiki M, Galanis D, Pavlopoulos J, Papageorgiou H, Androutsopoulos I, Manandhar S (2014) Semeval-2014 task 4: aspect based sentiment analysis, pp 27–35. https://doi.org/10.3115/v1/s14-2004
57. Pontiki M, Galanis D, Papageorgiou H, Androutsopoulos I, Manandhar S, Al-Smadi M, Al-Ayyoub M, Zhao Y, Qin B, Clercq OD, Hoste V, Apidianaki M, Tannier X, Loukachevitch N.V, Kotelnikov E.V, Bel N, Zafra S.M.J, Eryigit G (2016) Semeval-2016 task 5: Aspect based sentiment analysis, pp 19–30. https://doi.org/10.18653/v1/s16-1002
58. Zhang C, Li Q, Song D (2019) Aspect-based sentiment classification with aspect-specific graph convolutional networks, pp 4567–4577. https://doi.org/10.18653/v1/D19-1464
59. Liu Q, Zhang H, Zeng Y, Huang Z, Wu Z (2018) Content attention model for aspect based sentiment analysis, pp 1023–1032. https://doi.org/10.1145/3178876.3186001
60. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
61. Wang K, Shen W, Yang Y, Quan X, Wang R (2020) Relational graph attention network for aspect-based sentiment analysis, pp 3229–3238 .https://doi.org/10.18653/V1/2020.ACL-MAIN.295
62. Morris JX, Lifland E, Yoo JY, Grigsby J, Jin D, Qi Y (2020) Textattack: a framework for adversarial attacks, data augmentation, and adversarial training in NLP, pp 119–126. https://doi.org/10.18653/v1/2020.emnlp-demos.16
63. Mosbach M, Andriushchenko M, Klakow D (2021) On the stability of fine-tuning BERT: misconceptions, explanations, and strong baselines. The International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=nzpLWnVAyah
64. Liu X, Zhong Y, Wang J, Li P (2023) Data augmentation using heuristic masked language modeling. Int J Mach Learn Cybernet 14:2591–26050
65. Chen Z, Qian T (2022) Description and demonstration guided data augmentation for sequence tagging. World Wide Web 25(1):175–194. https://doi.org/10.1007/s11280-021-00978-0
66. Kolomiyets O, Bethard S, Moens M (2011) Model-portability experiments for textual temporal analysis. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, Assoc Comput Linguist. Portland, Oregon, USA, 271–276

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.