# An Empirical Study of Sentiment-Enhanced Pre-Training for Aspect-Based Sentiment Analysis

**Yice Zhang**[1,2], **Yifan Yang**[1,3], **Bin Liang**[1,3], **Shiwei Chen**[1,2], **Bing Qin**[1,2], **and Ruifeng Xu**[1,2,3*]

[1] Harbin Insitute of Technology, Shenzhen, China

[2] Peng Cheng Laboratory, Shenzhen, China

[3] Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies

{zhangyc_hit,evanyfyang}@163.com

bin.liang@stu.hit.edu.cn, chenshw@pcl.ac.cn

qinb@ir.hit.ac.cn, xuruifeng@hit.edu.cn

## Abstract

Aspect-Based Sentiment Analysis (ABSA) aims to recognize fine-grained opinions and sentiments of users, which is an important problem in sentiment analysis. Recent work has shown that Sentiment-enhanced Pre-Training (SPT) can substantially improve the performance of various ABSA tasks. However, there is currently a lack of comprehensive evaluation and fair comparison of existing SPT approaches. Therefore, this paper performs an empirical study to investigate the effectiveness of different SPT approaches. First, we develop an effective knowledge-mining method and leverage it to build a large-scale knowledge-annotated SPT corpus. Second, we systematically analyze the impact of integrating sentiment knowledge and other linguistic knowledge in pre-training. For each type of sentiment knowledge, we also examine and compare multiple integration methods. Finally, we conduct extensive experiments on a wide range of ABSA tasks to see how much SPT can facilitate the understanding of aspect-level sentiments.[1]

## 1 Introduction

Aspect-Based Sentiment Analysis (ABSA) is an important problem in sentiment analysis (Pontiki et al., 2014). Its goal is to recognize opinions and sentiments towards specific aspects from user-generated content (Zhang et al., 2022). Traditional ABSA approaches generally develop several separate models (Xu et al., 2018; Xue and Li, 2018; Fan et al., 2019) or a joint model (He et al., 2019; Chen and Qian, 2020), establishing interactions between different sentiment elements through specific model structures.

In recent years, pre-trained models (Devlin et al., 2019; Liu et al., 2019; Raffel et al., 2020) have yielded excellent results in extensive NLP tasks. This inspires many research efforts that leverage pre-training techniques to learn sentiment-aware representations. Among them, Xu et al. (2019a) reveal that pre-training on the sentiment-dense corpus through masked language modeling alone could result in significant improvements on three downstream ABSA tasks. Further, researchers undertake many explorations on integrating sentiment knowledge (*e.g.*, sentiment words) in the pre-training phase (Tian et al., 2020; Zhou et al., 2020; Ke et al., 2020; Li et al., 2021; Fan et al., 2022), as sentiment knowledge has been widely demonstrated to be helpful in various ABSA tasks (Li and Lam, 2017; Zeng et al., 2019; He et al., 2019; Xu et al., 2020a; Wu et al., 2020b; Liang et al., 2022).

Despite significant gains in various ABSA tasks, there has not been a comprehensive evaluation and fair comparison of existing Sentiment-enhanced Pre-Training (SPT) approaches. Therefore, this paper conducts an empirical study of SPT-ABSA to systematically investigate and analyze the effectiveness of the existing approaches. We mainly concentrate on the following questions: (a) what impact do different types of sentiment knowledge have on downstream ABSA tasks?; (b) which knowledge integration method is most effective?; and (c) does injecting non-sentiment-specific linguistic knowledge (*e.g.*, part-of-speech tags and syntactic relations) into pre-training have positive impacts? Based on the experimental investigation of these questions, we eventually obtain a powerful sentiment-enhanced pre-trained model. We evaluate it on a wide range of ABSA tasks to see how much SPT can facilitate the understanding of aspect-level sentiments.

* Corresponding Authors

[1] We release our code, data, and pre-trained model weights at https://github.com/HITSZ-HLT/SPT-ABSA.



*Findings of the Association for Computational Linguistics: ACL 2023*, pages 9633–9651

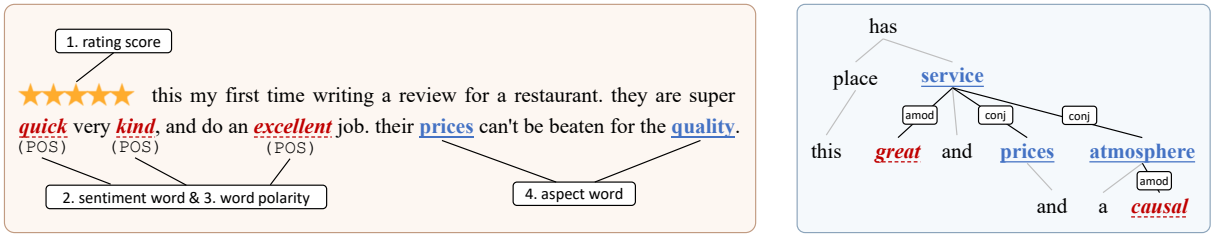


Figure 1: (a) left: four types of sentiment knowledge. (b) right: dependency links between aspect words and sentiment words. Aspect words and sentiment words are marked with blue and orange, respectively.

To enable our study, we prepare a large-scale knowledge-annotated SPT corpus. We obtain and collate over 100 million user-generated reviews from Yelp and Amazon. Subsequently, we develop an effective semi-supervised method for sentiment knowledge mining and annotating. This method is driven by lexicons and syntactic rules, and we devise an Expectation-Maximization (EM) algorithm to estimate them. Experiments demonstrate that this method can mine more considerable and accurate sentiment knowledge than the existing methods.

Our contributions can be concluded as follows:

- We develop an effective sentiment knowledge mining method and leverage it to build a large-scale knowledge-annotated SPT corpus.
- We systematically review and summarize the existing SPT approaches and empirically investigate and analyze their effectiveness.
- We conduct extensive experiments on ABSA tasks and illustrate how SPT can facilitate the understanding of aspect-level sentiments.

## 2 Analysis Setup

### 2.1 Pre-training Data

Following Xu et al. (2019a), we use user-generated reviews from Yelp datasets[2] and Amazon reviews datasets[3] (Ni et al., 2019) for pre-training. We remove those reviews that are too short (<50 characters) and too long (>500 characters) and end up with a corpus containing 140 million reviews in 28 domains. Its statistic is detailed in Appendix A.1.

#### 2.1.1 Sentiment Knowledge Mining

In this paper, we mainly investigate four typical types of sentiment knowledge: reviews' rating scores, sentiment words, word sentiment polarity, and aspect words. We illustrate them in Figure 1(a). Since only annotations of rating scores exist in the collected pre-training corpus, we develop an effective semi-supervised sentiment knowledge mining method.

Our method draws inspiration from the double propagation algorithm proposed by Qiu et al. (2011). They observe that there are some syntactic patterns linking aspect words and sentiment words, which is illustrated in Figure 1(b). Consequently, they define some syntactic rules to expand the aspect lexicon and sentiment lexicon iteratively. However, their method requires careful manual selection of syntactic rules. This limitation hinders the exploitation of complex syntactic patterns, such as (*pizza*, *awful*) in "*we had a lamb pie pizza that was awful*".

To overcome the above limitation, we devise an Expectation-Maximization (EM) algorithm to learn syntactic rules. In our method, the annotations of sentiment words and aspect words in the reviews are treated as unobserved latent variables, and the lexicons and rules are treated as the parameters. We first initialize parameters using MPQA (Wilson et al., 2005) and several simple syntactic rules; E-step annotates the reviews through the current estimate for the parameters; M-step updates the parameters according to the expected annotations. This process can be formulated as:

$$\text{initialize } \boldsymbol{\theta}^{(0)}, \tag{1}$$

repeat:

$$\boldsymbol{y}_{\mathcal{S}}^{(t)}, \boldsymbol{y}_{\mathcal{A}}^{(t)} = \mathbb{E}(\boldsymbol{x}; \boldsymbol{\theta}^{(t)}), \tag{2}$$

$$\boldsymbol{\theta}^{(t+1)} = \mathbb{M}(\boldsymbol{x}, \boldsymbol{y}_{\mathcal{S}}^{(t)}, \boldsymbol{y}_{\mathcal{A}}^{(t)}), \tag{3}$$

where $\boldsymbol{\theta} = (\mathcal{L}_{\mathcal{S}}, \mathcal{L}_{\mathcal{A}}, \mathcal{P}_{\mathcal{SS}}, \mathcal{P}_{\mathcal{AA}}, \mathcal{P}_{\mathcal{SA}}, \mathcal{P}_{\mathcal{AS}})$ denotes the lexicons and syntactic rules. For each mined sentiment word, we use Pointwise Mutual Information (PMI) to determine its polarity (Turney, 2002; Tian et al., 2020). See Appendix B for

[2] https://www.yelp.com/dataset

[3] https://nijianmo.github.io/amazon/index.html

| Task | Input | Output |
|---|---|---|
| AE | [CLS] *delicious mushroom pizza but slow and rude delivery* [SEP] | S O B I O O O O B E |
| ASC | [CLS] *delicious mushroom pizza but slow and rude delivery* [SEP] *mushroom pizza* [SEP] | POS |
| | [CLS] *delicious mushroom pizza but slow and rude delivery* [SEP] *delivery* [SEP] | NEG |
| AOE | [CLS] *delicious mushroom pizza but slow and rude delivery* [SEP] *mushroom pizza* [SEP] | S B O O O O O O O O O O E |
| | [CLS] *delicious mushroom pizza but slow and rude delivery* [SEP] *delivery* [SEP] | S O O O O B O B O O O E |

Table 1: Examples of three downstream ABSA tasks.

more details.

#### 2.1.2 Syntax Knowledge Acquisition

We annotate four types of syntax knowledge in the reviews using `spaCy`[4]. For each word, we annotate its *part-of-speech tag*. If there is a dependency relation between two words, we annotate its *direction* and *type*. If a word is the ancestor of another word, we annotate their *dependency distance*.

### 2.2 Downstream Tasks and Datasets

An aspect-level opinion can be defined as a triplet consisting of an aspect term, the corresponding opinion term, and the sentiment polarity (Peng et al., 2020). Therefore, we select Aspect term Extraction (AE), Aspect-oriented Opinion term Extraction (AOE), and Aspect-level Sentiment Classification (ASC) to measure a model's understanding of these three sentiment elements, respectively. These downstream tasks are illustrated in Table 1. The datasets for these three ABSA tasks are derived from Wang et al. (2017); Fan et al. (2019). Their statistics are detailed in Appendix A.2.

## 3 Method

Given a review $\boldsymbol{X}$ of length $T$, a pre-trained model produces its word-level contextualized representations and review-level representation, which can be generally formulated as follows:

$$\boldsymbol{h}_{[CLS]}, \boldsymbol{h}_1, \cdots, \boldsymbol{h}_T = f(x_1, \cdots, x_T; \boldsymbol{\theta}_{PLM}).$$

General-purpose pre-training (Devlin et al., 2019; Liu et al., 2019; Lan et al., 2020; He et al., 2020) mostly learns parameters through Masked Language Modeling (MLM). In MLM, a certain proportion of words $\mathcal{C}$ in the review is masked, and the masked review $\tilde{\boldsymbol{X}}$ is then input to the pre-trained model to recover the masked part:

$$P(\tilde{x}_t) = \text{softmax}(\text{FFNN}(\boldsymbol{h}_t)), \quad (4)$$

$$\mathcal{L}_{MLM} = -\frac{1}{|\mathcal{C}|} \sum_{t \in \mathcal{C}} \log P(\tilde{x}_t = x_t), \quad (5)$$

where FFNN denotes a feed-forward neural network with non-linear activation, and for simplicity, we still use $\boldsymbol{h}_t$ to denote the word-level representation of $\tilde{\boldsymbol{X}}$.

Existing SPT approaches integrate sentiment knowledge in two main ways: (1) **knowledge-guided masking** prioritizes masking sentiment knowledge in reviews and leverages MLM to increase the model's awareness of sentiment knowledge; (2) **knowledge supervision** directly converts sentiment knowledge into labels and then predicts them by the word-level representations and the review-level representation.

### 3.1 Integrating Aspect & Sentiment Words

A common way to integrate aspect and sentiment words is to increase their masking probabilities (Tian et al., 2020; Zhou et al., 2020; Ke et al., 2020; Li et al., 2021). There are two implementations: (1) **mask-by-probability** masks these words with probability $x\%$ and masks other words with probability 15%; (2) **mask-by-proportion** randomly masks these words to $y\%$ of the total words and masks other words to $(15-y)\%$ of the total words. The main difference between these two implementations is that the former is more sensitive to the proportion of aspect and sentiment words in the review.

In addition to increasing their masking probability, we propose the strategy of masking their contexts. Our motivation stems from the observation sentiment expressions are often closer to aspect words, and thus we can leverage aspect words to locate sentiment-dense segments of a review. We assign a higher masking probability to words that are closer to aspect words. To achieve this, we

[4]The trained pipeline we use is `en_core_web_sm 3.3.0`.

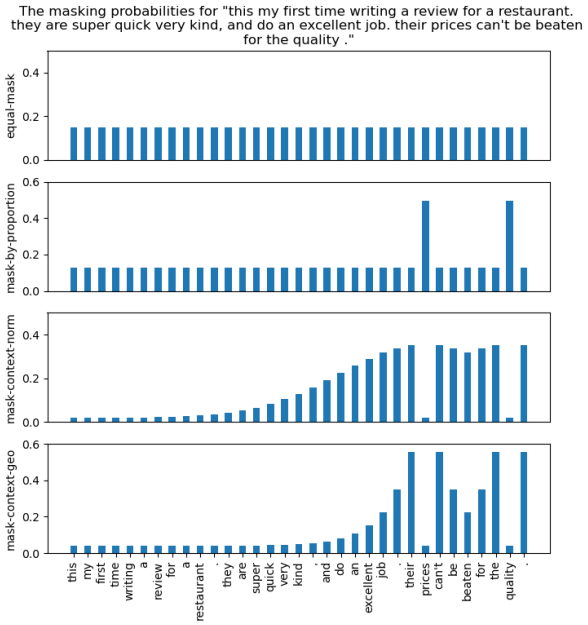


Figure 2: Illustration of masking strategies for integrating aspect words.

empirically choose the normal distribution and geometric distribution for the masking probability assignment, and the corresponding masking strategies are denoted as **mask-context-norm** and **mask-context-geo**. Figure 2 provides an illustration of these two masking strategies, and detailed implementations can be found in Appendix C.

Moreover, an alternative way is to convert the aspect and sentiment words to **pseudo-label**s and then use the word-level representations to predict these pseudo-labels, which can be formulated as:

$$P(y_t) = \text{softmax}(\text{FFNN}(\boldsymbol{h}_t)), \quad (6)$$

$$\mathcal{L}_{\mathcal{A}/\mathcal{S}\text{-}PL} = -\frac{1}{T}\sum_t \log P(y_t^*), \quad (7)$$

where $y_t \in \{\texttt{AspW}, \texttt{Other}\}$ for integrating aspect words, and $y_t \in \{\texttt{SenW}, \texttt{Other}\}$ for integrating sentiment words.

### 3.2 Integrating Review Rating

The review's rating score reflects its overall sentiment. To integrate it, Zhou et al. (2020); Ke et al. (2020) introduce rating prediction. They predict the rating score by the review-level representation and use the **cross-entropy** function to calculate the loss:

$$P(y_{RAT}) = \text{softmax}(\text{Linear}(\boldsymbol{h}_{[CLS]})), \quad (8)$$

$$\mathcal{L}_{RAT\text{-}CE} = -\log P(y_{RAT}^*), \quad (9)$$

where $y_{RAT} \in \{\texttt{1}, \texttt{2}, \texttt{3}, \texttt{4}, \texttt{5}\}$.

Besides, Li et al. (2021) adopt the **supervised-contrastive-learning** objective (Khosla et al., 2020) to integrate review rating. With this objective, representations from the same sentiment are pulled closer together than representations from different sentiments. Specifically, the loss for a batch $\mathcal{B}$ is calculated as follows:

$$\ell(i,j) = -\log \frac{\exp(\text{sim}(i,j)/\tau)}{\sum_{k\in\mathcal{B}\setminus i}\exp(\text{sim}(i,k)/\tau)}, \quad (10)$$

$$\mathcal{L}_{RAT\text{-}SCL} = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\frac{1}{|P(i)|}\sum_{j\in P(i)}\ell(i,j), \quad (11)$$

where $P(i) = \{j | j \in \mathcal{B}\setminus i, y_{RAT}^{(i)} = y_{RAT}^{(j)}\}$ is the set of indices of all positives in batch $\mathcal{B}$ for $i$, and $\tau$ is the temperature. They use $\text{sim}(i,j) = \boldsymbol{s}_i^\top \boldsymbol{s}_j$ on the normalized representations $\boldsymbol{s}_i = \boldsymbol{h}_{[CLS]}^{(i)}/\|\boldsymbol{h}_{[CLS]}^{(i)}\|$ as similarity metric.

### 3.3 Integrating Other Sentiment Knowledge

**Word Polarity**. To integrate this knowledge, Tian et al. (2020); Zhou et al. (2020); Ke et al. (2020) introduce the objective of word polarity prediction. In this objective, word polarity is inferred based on word-level representations, similar to Equation 6. There are two variants: one only predicts the polarity of **masked-sentiment-words**, *i.e.*, $y_t \in \{\texttt{POS}, \texttt{NEG}\}$; the other predicts the polarity of **all-masked-words**, *i.e.*, $y_t \in \{\texttt{POS}, \texttt{NEG}, \texttt{Other}\}$. The difference is that the latter also includes the label of sentiment words in the supervision.

**Aspect-Sentiment Pair.** Tian et al. (2020) regard a sentiment word with its nearest noun (the maximum distance is 3) as an aspect-sentiment pair. They argue that aspect-sentiment pairs reveal more information than sentiment words do. Therefore, they propose aspect-sentiment pair prediction to capture the dependency between aspect and sentiment. They randomly mask at most 2 aspect-sentiment pairs in each review and predict them through the review-level representation[5]:

$$P(x) = \text{softmax}(\text{FFNN}(\boldsymbol{h}_{[CLS]})), \quad (12)$$

$$\mathcal{L}_{PAIR} = -\frac{1}{|\mathcal{P}|}\sum_{t\in\mathcal{P}}\log P(x = x_t), \quad (13)$$

where $\mathcal{P}$ is the set of indices of words in the masked aspect-sentiment pairs.

[5]There are some differences in the loss calculation from the original paper (Tian et al., 2020). See our note on this in Appendix D.

| Knowledge | Method | Restaurant-14 | | | Laptop-14 | | | Avg-Δ |
|---|---|---|---|---|---|---|---|---|
| | | AE | ASC | AOE | AE | ASC | AOE | |
| BERT | | 86.41 | 75.83 | 85.44 | 80.49 | 72.85 | 78.10 | - |
| MLM | | 88.26(+1.85) | 78.63(+2.80) | 86.04(+0.60) | 81.55(+1.06) | 73.38(+0.53) | 79.90(+1.80) | +1.44 |
| +AspectWord | mask-by-probability | 87.93(+1.52) | 78.82(+2.99) | 86.04(+0.60) | 81.47(+0.98) | 73.83(+0.98) | 79.61(+1.51) | +1.43 |
| | mask-by-proportion | 87.35(+0.94) | 78.54(+2.71) | 86.06(+0.62) | 81.76(+1.27) | 73.51(+0.66) | 79.66(+1.56) | +1.29 |
| | mask-context-norm | 88.30(+1.89) | 79.13(+3.30) | 85.77(+0.33) | 81.66(+1.17) | 74.48(+1.63) | 79.90(+1.80) | +1.69 |
| | ★ mask-context-geo | 88.30(+1.89) | 79.04(+3.21) | 86.38(+0.94) | 81.80(+1.31) | 73.99(+1.14) | 80.40(+2.30) | **+1.80** |
| | pseudo-label | 88.25(+1.84) | 78.81(+2.98) | 86.11(+0.67) | 82.04(+1.55) | 73.41(+0.56) | 80.11(+2.01) | +1.61 |
| +SentimentWord | mask-by-probability | 88.47(+2.06) | 78.92(+3.09) | 85.94(+0.50) | 81.24(+0.75) | 74.28(+1.43) | 79.76(+1.66) | +1.59 |
| | ★ mask-by-proportion | 88.17(+1.76) | **79.34**(+3.51) | 85.83(+0.39) | 81.16(+0.67) | **75.71**(+2.86) | 79.44(+1.34) | **+1.76** |
| | pseudo-label | 88.35(+1.94) | 78.61(+2.78) | 85.70(+0.26) | 81.33(+0.84) | 74.06(+1.21) | 80.39(+2.29) | +1.56 |
| +ReviewRating | ★ cross-entropy | 87.95(+1.54) | **80.02**(+4.19) | 85.93(+0.49) | 80.89(+0.40) | **75.78**(+2.93) | 79.92(+1.82) | **+1.90** |
| | contrastive-learning | 88.27(+1.86) | **79.26**(+3.43) | 86.37(+0.93) | 81.10(+0.61) | **75.70**(+2.85) | 80.05(+1.95) | **+1.94** |
| +WordPolarity | masked-sentiment-words | 88.18(+1.77) | 78.61(+2.78) | 86.03(+0.59) | 80.57(+0.08) | 74.02(+1.17) | 79.95(+1.85) | +1.38 |
| | ★ all-masked-words | 88.14(+1.73) | **79.50**(+3.67) | 85.90(+0.46) | 81.41(+0.92) | 73.86(+1.01) | 79.87(+1.77) | +1.60 |
| +Aspect-Sentiment-Pair | | 87.74(+1.33) | **80.14**(+4.31) | 86.30(+0.86) | 81.75(+1.26) | **75.77**(+2.92) | 78.93(+0.83) | **+1.92** |
| +Emoticon | | 88.33(+1.92) | 79.02(+3.19) | 86.27(+0.83) | 81.17(+0.68) | 73.29(+0.44) | 79.64(+1.54) | +1.44 |
| +Syntax | part-of-speech (pos) | **89.00**(+2.59) | 78.08(+2.25) | 86.14(+0.70) | **82.02**(+1.53) | 73.76(+0.91) | 80.12(+2.02) | +1.67 |
| | dependency-direction | **88.73**(+2.32) | 78.72(+2.89) | **86.54**(+1.10) | 81.68(+1.19) | 73.62(+0.77) | 79.91(+1.81) | +1.68 |
| | dependency-type | 88.47(+2.06) | 78.04(+2.21) | 86.15(+0.71) | 81.86(+1.37) | 74.46(+1.61) | 80.12(+2.02) | +1.67 |
| | dependency-distance | 88.44(+2.03) | 78.09(+2.26) | **86.80**(+1.36) | 81.85(+1.36) | 73.99(+1.14) | **80.54**(+2.44) | **+1.77** |
| | ★ pos & direction & distance | **88.78**(+2.37) | 78.03(+2.20) | **86.55**(+1.11) | 81.88(+1.39) | 73.77(+0.92) | **80.89**(+2.79) | **+1.80** |
| +AspectWord+SentimentWord | | 88.29(+1.88) | 78.98(+3.15) | 85.99(+0.55) | 81.37(+0.88) | 74.14(+1.29) | 80.11(+2.01) | +1.63 |
| +AspectWord+Rating | | 88.18(+1.77) | **79.98**(+4.15) | 86.21(+0.77) | 81.48(+0.99) | **76.00**(+3.15) | 79.54(+1.44) | **+2.05** |
| +AspectWord+Pair | | 87.75(+1.34) | **79.83**(+4.00) | 86.00(+0.56) | 81.49(+1.00) | **75.49**(+2.64) | 78.80(+0.70) | +1.71 |
| +AspectWord+Syntax | | **88.71**(+2.30) | 78.29(+2.46) | **86.75**(+1.31) | **82.33**(+1.84) | 74.00(+1.15) | **80.72**(+2.62) | **+1.95** |
| +Rating+Syntax | | **88.70**(+2.29) | **79.81**(+3.98) | 86.17(+0.73) | **82.18**(+1.69) | **76.13**(+3.28) | 80.32(+2.22) | **+2.37** |
| ★ +AspectWord+Rating+Syntax | | **88.66**(+2.25) | **79.88**(+4.05) | 86.31(+0.87) | **82.22**(+1.73) | **76.42**(+3.57) | 80.22(+2.12) | **+2.42** |

Table 2: Performance of integrating different knowledge in pre-training ($F_1$-score, %). The evaluation metric for ASC is Macro-$F_1$. We boldface those results with significant advantages. For each type of knowledge, we mark the best integration approach with a ★. Among the two methods of integrating review ratings, although supervised contrastive learning performs slightly better than cross-entropy, the latter is simpler and more straightforward. Therefore, we mark cross-entropy as the preferred method.

**Emoticon**s (*e.g.*, ":-)" and ":-(") are often inserted in the reviews to express emotions. Zhou et al. (2020) point out that integrating emoticons can capture more token-level sentiment knowledge. Consistent with Zhou et al. (2020), we treat emoticons as special tokens during the tokenization process and assign them a masking probability of 50% when masking.

### 3.4 Integrating Syntax Knowledge

Although syntax knowledge has been widely incorporated in fine-tuning various ABSA tasks (Zhang et al., 2019; Huang and Carley, 2019; Wang et al., 2020; Chen et al., 2022), few works explore its impact on SPT. In this paper, we cover four types of syntax knowledge and integrate them through knowledge supervision. We infer **part-of-speech** tags in the same way as in Equation 6 and transform the predictions of **dependency-direction**, **dependency-type**, and **dependency-distance** into word-pair classification. The word-pair classification can be formulated as follows:

$$\tilde{\boldsymbol{h}}_i = \text{FFNN}(\boldsymbol{h}_i), \tilde{\boldsymbol{h}}_j = \text{FFNN}(\boldsymbol{h}_j), \quad (14)$$

$$\tilde{\boldsymbol{h}}_{i\to j} = [\tilde{\boldsymbol{h}}_i; \tilde{\boldsymbol{h}}_j; \tilde{\boldsymbol{h}}_i - \tilde{\boldsymbol{h}}_j; \tilde{\boldsymbol{h}}_i * \tilde{\boldsymbol{h}}_j], \quad (15)$$

$$P(y_{i\to j}) = \text{softmax}(\text{Linear}(\tilde{\boldsymbol{h}}_{i\to j})), \quad (16)$$

$$\mathcal{L}_{DIR/TYP/DIS} = -\sum_{i,j} \log P(y^*_{i\to j}). \quad (17)$$

## 4 Experiment

### 4.1 Implementation Details

Following Xu et al. (2019a); Zhou et al. (2020); Li et al. (2021), we use BERT (Devlin et al., 2019) as the base framework and initialize the model

| Backbone | Step | Restaurant-14 | | | Laptop-14 | | | Avg. |
|---|---|---|---|---|---|---|---|---|
| | | AE | ASC | AOE | AE | ASC | AOE | |
| BERT | - | 86.41 | 75.83 | 85.44 | 80.49 | 72.85 | 78.10 | 79.86 |
| +SKEP*(Tian et al., 2020) | 10k | 87.09 | 79.85 | 85.15 | 80.29 | 75.56 | 79.03 | 81.17(+1.31) |
| +SentiLARE*(Ke et al., 2020) | 10k | 88.27 | 79.29 | 86.17 | 81.72 | 75.54 | 80.60 | 81.93(+2.07) |
| **+Our SPT** | 10k | **88.66** | **79.88** | **86.31** | **82.22** | **76.42** | 80.22 | **82.28(+2.42)** |
| +SCAPT$_{\text{LAP}}$(Li et al., 2021) | 24k | 86.38 | 78.99 | 86.54 | 83.09 | 75.69 | 79.56 | 81.71(+1.85) |
| +SCAPT$_{\text{REST}}$(Li et al., 2021) | 75k | 87.89 | 79.01 | 86.10 | 80.30 | 76.60 | 78.83 | 81.46(+1.60) |
| **+Our SPT** | 24k | **88.79** | **79.27** | 86.22 | 82.67 | **77.31** | **80.84** | **82.52(+2.66)** |
| +SentiX(Zhou et al., 2020) | 280k | 87.08 | 78.53 | 85.37 | 80.50 | 75.36 | 78.95 | 80.97(+1.11) |
| **+Our SPT** | 280k | **88.54** | **81.10** | **86.51** | **83.39** | **77.70** | **79.39** | **82.77(+2.81)** |
| +BERT$_{\text{REVIEW}}$(Xu et al., 2019b) | 800k | 89.12 | 78.99 | 86.23 | 84.32 | 75.94 | 80.72 | 82.55(+2.69) |
| **+Our SPT** | 400k | 88.91 | **81.59** | **86.79** | 83.83 | **78.68** | 78.88 | **83.11(+3.25)** |

Table 3: Comparison results with the previous SPT works. Our SPT refers to the combination of aspect words, review ratings, and syntax knowledge. The original SKEP and SentiLARE are not pre-trained based on `BERT-uncased-base`, so we reproduce them on our SPT corpus. We convert the computational cost into training steps. See our notes in Appendix D for this conversion.

weights through `BERT-base-uncased`. We implement pre-training with a batch size of 1000, and an initial learning rate of 2e-4. See Appendix E for the detailed hyper-parameters. Our pre-training corpus covers 28 domains, such as `Restaurant`, `Laptop`, and `Books`. For most experiments, we only pre-train 10k steps on both `Restaurant` and `Laptop`. To mitigate the effect of randomness, we run each pre-training approach twice, evaluate each pre-trained model on three downstream tasks 10 times, and release the average results. Moreover, we also pre-train 400k steps on all domains to fully exploit the potential of SPT.

### 4.2 Main Results

We continue to pre-train BERT via the different SPT approaches and subsequently fine-tune them on three ABSA tasks. Their performance is reported in Table 2. We see that MLM alone yields notable improvements, where the maximum is achieved on the ASC task in `Restaurant-14`, nearly 3%. Integrating sentiment and syntax knowledge leads to a variety of impacts.

**What impact do different types of sentiment knowledge have on downstream ABSA tasks?** Most sentiment knowledge contributes to performance improvement on the ASC task, with sentiment words, review ratings, and aspect-sentiment pairs showing the highest potential. Integrating aspect words provides general benefits, and masking their contexts improves performance on nearly all downstream tasks. The impact of integrating emoticons is minimal.

**Which knowledge integration method is most effective?** For aspect words, masking their contexts has a generally positive impact, while increasing their masking probabilities does not. This finding suggests that predicting their context is helpful as the context often contains the key cues of its sentiment. For sentiment words, mask-by-proportion is better than mask-by-probability. This is because the former can better balance the masking proportion of sentiment knowledge and general knowledge. For review ratings, we observe that using supervised contrastive learning does not show a salient advantage over cross-entropy, indicating that the application of contrastive learning on SPT still needs exploration.

**Does integrating syntax knowledge have positive impacts?** Aspect terms are often phrases, such as *the orecchiette with sausage and chicken*. These phrases tend to follow certain part-of-speech patterns. From Table 2, we observe that integrating part-of-speech helps the model extract aspect terms. Furthermore, Pouran Ben Veyseh et al. (2020) state that the dependency structure can provide useful information to improve the performance of AOE. Our experimental results align with their statement.

**Whether integrating multiple knowledge simultaneously can lead to better results?** According to the results in Table 2, we find that integrating multiple knowledge simultaneously does not necessarily lead to better performance. The most

| | Backbone | Rest14 | Lap14 | Rest15 | Rest16 | Avg-Δ |
|---|---|---|---|---|---|---|
| GTS | BERT | 68.52 | 55.26 | 59.70 | 67.08 | - |
| | +SKEP* | 71.86 | 56.89 | 63.44 | 69.72 | +2.84 |
| | +SentiLARE* | 70.99 | 56.82 | 64.57 | 69.99 | +2.60 |
| | +SCAPT$_{\text{LAP}}$ | 69.76 | 58.73 | 60.87 | 68.15 | +1.74 |
| | +SCAPT$_{\text{REST}}$ | 71.92 | 55.67 | 63.37 | 68.79 | +2.30 |
| | +SentiX | 71.37 | 58.50 | 63.12 | 70.71 | +3.29 |
| | +BERT$_{\text{REVIEW}}$ | 71.40 | 58.42 | 63.46 | 68.49 | +2.80 |
| | **+Our SPT (10k)** | 71.44 | 56.19 | 64.77 | 70.63 | +3.12 |
| | **+Our SPT (400k)** | **72.11** | 58.27 | **64.98** | 68.44 | **+3.31** |
| BMRC | BERT | 70.32 | 58.96 | 60.70 | 67.30 | - |
| | +SKEP* | 71.16 | 59.46 | 63.08 | 68.84 | +1.32 |
| | +SentiLARE* | 71.40 | 60.70 | 62.86 | 69.51 | +1.80 |
| | +SCAPT$_{\text{LAP}}$ | 70.99 | 59.51 | 59.15 | 67.33 | -0.08 |
| | +SCAPT$_{\text{REST}}$ | 71.17 | 57.22 | 61.26 | 67.55 | -0.02 |
| | +SentiX | 69.54 | 60.42 | 60.71 | 67.79 | +0.30 |
| | +BERT$_{\text{REVIEW}}$ | 71.25 | 62.20 | 64.50 | 70.68 | +2.84 |
| | **+Our SPT (10k)** | 70.83 | 59.47 | 63.03 | 69.62 | +1.42 |
| | **+Our SPT (400k)** | **71.43** | **62.40** | 64.32 | 70.46 | +2.83 |
| Span-ASTE | BERT | 72.25 | 59.45 | 63.58 | 70.26 | - |
| | +SKEP* | 73.47 | 62.07 | 65.32 | 72.51 | +1.98 |
| | +SentiLARE* | 74.71 | 62.86 | 66.13 | 73.19 | +2.86 |
| | +SCAPT$_{\text{LAP}}$ | 72.60 | 61.62 | 61.00 | 70.69 | +0.11 |
| | +SCAPT$_{\text{REST}}$ | 74.23 | 59.01 | 64.08 | 71.16 | +0.76 |
| | +SentiX | 72.69 | 62.25 | 64.26 | 71.44 | +1.30 |
| | +BERT$_{\text{REVIEW}}$ | 74.31 | 63.51 | 64.23 | 72.55 | +2.29 |
| | **+Our SPT (10k)** | 74.74 | 62.22 | 67.55 | **73.72** | +3.19 |
| | **+Our SPT (400k)** | **75.34** | **64.76** | **67.69** | 73.49 | **+3.93** |

Table 4: Performance on the ASTE task ($F_1$-score, %). Results are the average of 5 runs.

obvious evidence lies in the combination of aspect words and aspect-sentiment pairs. In most scenarios, the introduction of an additional type of knowledge brings both benefits and drawbacks. Experimental results highlight that the combination of aspect words, review ratings, and syntax knowledge achieves the best trade-off, yielding an average improvement of 2.42% over BERT.

Further, we compare the best combination with previous SPT works and present the results in Table 3. These results show that under the same computational cost, this combination outperforms previous works in most cases. When pre-training 400k steps, we observe an average improvement of 3.25% over BERT. In addition, we note an anomaly in the model pre-trained on all domains. Specifically, its performance on the AOE task in `Laptop-14` does not increase with the number of pre-training steps. This may be because the dependency between the aspect term and opinion term varies between domains. Further exploration of this phenomenon is warranted in future research.

| Backbone | AE | ABSA |
|---|---|---|
| Previous SOTA | 50.00 | 43.46 |
| BERT | 38.99 | 36.13 |
| +SKEP* | 42.37 | 40.43 |
| +SentiLARE* | 42.19 | 41.22 |
| +SCAPT$_{\text{LAP}}$ | 42.21 | 39.90 |
| +SCAPT$_{\text{REST}}$ | 41.07 | 39.11 |
| +SentiX | 42.70 | 40.76 |
| +BERT$_{\text{REVIEW}}$ | 45.57 | 43.77 |
| **+Our SPT (10k)** | 43.09 | 41.59 |
| +AspectWord+Rating (10k) | 42.49 | 40.75 |
| **+Our SPT (400k)** | **45.61** | **44.16** |

Table 5: Performance on two cross-domain ABSA tasks ($F_1$-score, %). This table only presents the average performance of AE and end-to-end ABSA, and full results are listed in Appendix F. Performance of the previous SOTA comes from Yu et al. (2021). AspectWord+Rating (10k) denotes removing syntax knowledge from Our SPT.

### 4.3 Results on More Downstream Tasks

In addition to the three basic ABSA tasks, we further evaluate the SPT model on more ABSA tasks and datasets.

**Aspect Sentiment Triplet Extraction** (ASTE) aims to extract the aspect terms along with the corresponding opinion terms and the expressed sentiment (Peng et al., 2020). As a compound task, ASTE evaluates the model's understanding of aspect-level sentiments comprehensively. We take the pre-trained model as the language encoder and select three classical methods for triplet extraction: GTS (Wu et al., 2020a), BMRC (Chen et al., 2021), and Span-ASTE (Xu et al., 2021). We conduct experiments on ASTE-Data-v2 (Xu et al., 2020b) and present the results in Table 4.

Experimental results show that SPT can generally improve the performance of ASTE. The best pre-trained model is our SPT (400k), which achieves an average improvement of 3.31% and 3.93% on GTS and Span-ASTE, respectively. Additionally, we observe that SPT has a smaller improvement on BMRC, suggesting that the paradigm of machine reading comprehension relies more on the model's understanding of natural language statements than on the quality of the representations of partial words.

**Cross-domain ABSA** aims to transfer ABSA annotations from a resource-rich domain to a resource-poor domain (Gong et al., 2020). This task requires the model to possess the ability to learn domain-

| Backbone | ATSA | ACSA |
|---|---|---|
| BERT | 82.64 | 80.19 |
| +SKEP* | 83.45 | 81.15 |
| +SentiLARE* | 83.42 | 80.98 |
| +SCAPT$_{\text{REST}}$ | 83.21 | 81.26 |
| +SentiX | 82.32 | 81.37 |
| +BERT$_{\text{REVIEW}}$ | 83.02 | 79.87 |
| **+Our SPT (10k)** | **83.65** | 81.23 |
| **+Our SPT (400k)** | 83.41 | **81.52** |

Table 6: Performance on MAMS (Macro-$F_1$, %). Results are the average of 10 runs.

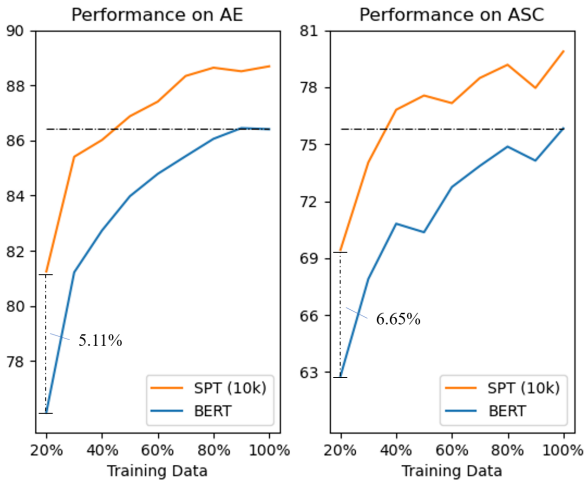


Figure 3: Performance of AE and ASC on Restaurant-14 under different amounts of training data.

| | Method | OE | | | AE | | |
|---|---|---|---|---|---|---|---|
| | | *P.* | *R.* | $F_1$ | *P.* | *R.* | $F_1$ |
| REST-14 | PoS | 52.70 | 51.39 | 52.04 | 67.54 | 22.13 | 33.34 |
| | Hu2004 | 78.26 | 70.24 | 74.03 | 69.02 | 35.36 | 46.76 |
| | MPQA | 76.49 | 73.51 | 74.97 | 66.67 | 35.98 | 46.74 |
| | **Ours** | **79.90** | **80.06** | **79.88** | **88.71** | **83.25** | **85.89** |
| LAP-14 | PoS | 38.49 | 44.81 | 41.41 | 43.43 | 17.74 | 25.19 |
| | Hu2004 | 63.62 | 56.82 | 60.03 | 34.50 | 17.74 | 23.43 |
| | MPQA | 69.48 | 65.88 | 67.63 | 41.96 | 19.27 | 26.41 |
| | **Ours** | **72.40** | **72.40** | **72.40** | **70.40** | **66.36** | **68.32** |

Table 7: Comparison results of knowledge-mining methods (Overlap-$F_1$, %). `PoS` denotes annotating sentiment words through pre-defined part-of-speech patterns (Tian et al., 2020). `MPQA` and `Hu2004` denotes annotating sentiment words through the sentiment lexicon provided by Deng and Wiebe (2015) and Hu and Liu (2004), respectively. Three baselines annotate the noun closest to every sentiment word as the aspect word.

invariant sentiment knowledge. We leverage this task to evaluate the cross-domain capabilities of pre-trained models. We conduct experiments on the datasets released by Gong et al. (2020) and list the results in Table 5.

We find that SPT greatly boosts the performance of BERT on cross-domain ABSA. The best models are those pre-trained on a mixture of multiple domains (BERT$_{\text{REVIEW}}$ and Our SPT (400k)). Besides, we notice that removing syntax knowledge causes a significant drop in performance, highlighting the importance of syntax knowledge in cross-domain ABSA. Despite achieving notable improvements over BERT, SPT alone has not yet achieved satisfactory performance, suggesting that addressing cross-domain ABSA requires more than just employing SPT.

**MAMS** (Jiang et al., 2019) is a challenging benchmark dataset for ABSA, where each review contains at least two different aspects with different sentiments. We list experimental results on MAMS in Table 6. We find that SPT still shows performance gains on this dataset, but only at most 1%, which is relatively lower than the gains observed on other datasets. This suggests that SPT for multi-aspect scenarios deserves further exploration.

### 4.4 Further Analysis

**Effect of SPT on Data-scarce Scenarios.** Data scarcity is a critical challenge in ABSA. We explore the effect of SPT under different amounts of training data. As depicted in Figure 3, the improvements from SPT become more obvious with less training data, with maximums of 5.11% and 6.65%. Furthermore, with SPT, the performance originally achieved using the entire training data can be attained using only 40% of it. This suggests that SPT is a feasible solution to alleviate the issue of data scarcity in ABSA.

**Evaluation for Knowledge-mining Methods**. We utilize Aspect term Extraction (AE) and Opinion term Extraction (OE) to indirectly evaluate knowledge-mining methods. Since opinion terms and aspect terms are typically phrases while mining results are at the word level, we use overlap-$F_1$ as the evaluation metric. The difference with the normal $F_1$-score is that overlap-$F_1$ recognizes a prediction as correct as long as it overlaps with any gold-truth term. We conduct experiments on the datasets provided by Wang et al. (2017) and list the results in Table 7. According to these results, our knowledge-mining method exhibits significant improvements over previous methods in both pre-

| Review Sentence | Ground-truth | BERT's Prediction | Our SPT(10k)'s Prediction |
|---|---|---|---|
| *cajun shrimp is good, not great* | NEU | POS ✘ | NEU ✔ |
| *i felt it was inferior in many ways to windows 7* | POS | NEG ✘ | POS ✔ |
| *screen - although some people might complain about low res* which i think is ridiculous | POS | NEG ✘ | POS ✔ |
| *his food is excellent (and not expensive by nyc standards- no entrees over 30, most appetizers 12 to 14)* | POS | NEU ✘ | NEU ✘ |
| *i opted for the squaretrade 3-year computer accidental protection warranty (1500-2000) which also support accidents like drops and spills that are not covered by applecare* | POS | NEU ✘ | NEU ✘ |

Table 8: Case Study on the ASC task. The aspect terms are marked with orange.

cision and recall. This superiority is particularly significant in AE. These observations demonstrate that our knowledge-mining method can mine more considerable and accurate sentiment knowledge.

**Case Study**. We present several representative examples in Table 8. The first three examples demonstrate that SPT can enable the model to discern sentiment polarities in complex semantics, such as comparisons and negations. However, the fourth and fifth examples highlight SPT's limitations in more intricate contexts, such as statements containing factual information. Consequently, more advanced techniques are required to enhance the model's understanding of sentiment.

## 5 Conclusion

In this paper, we perform an empirical study of Sentiment-enhanced Pre-Training (SPT). Our study investigates the impacts of integrating sentiment knowledge and other linguistic knowledge in pre-training on Aspect-Based Sentiment Analysis (ABSA). To enable our study, we first develop an effective knowledge-mining approach, leverage it to build a large-scale SPT corpus, and then select a range of ABSA tasks as the benchmark to systematically evaluate a pre-trained model's understanding of aspect-level sentiments. Experimental results reveal the following findings: (1) integrating aspect words brings general benefits to downstream tasks; (2) integrating sentiment words, review ratings, or aspect-sentiment pairs significantly improves the performance on aspect-level sentiment classification; (3) integrating syntax knowledge can help the model extract aspect terms and aspect-oriented opinion terms; and (4) the combination of aspect words, review ratings, and syntax knowledge achieves the best trade-off, yielding an average improvement of 3.25% over BERT. We further examine SPT's effectiveness on more ABSA tasks and find that SPT can improve the performance of a wide range of downstream tasks. Notably, SPT improves the model's cross-domain capabilities. In addition, we also demonstrate the effectiveness of our knowledge-mining method.

## Acknowledgments

We thank the anonymous reviewers for their valuable suggestions to improve the overall quality of this manuscript. This work was partially supported by the National Natural Science Foundation of China (62006062, 62176074, 62176076), Natural Science Foundation of Guangdong 2023A1515012922, the Shenzhen Foundational Research Funding (JCYJ20220818102415032, JCYJ20200109113441941), the Major Key Project of PCL2021A06, Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies 2022B1212010005, and Key Technologies Research and Development Program of Shenzhen JSGG20210802154400001.

## Limitations

While this paper provides a systematic investigation and analysis of existing Sentiment-enhanced Pre-Training (SPT) approaches, it is important to acknowledge the following limitations:

- The existing SPT approaches do not exhibit sufficient performance improvement in the multi-aspect scenario. Achieving significant performance enhancements in this particular scenario through SPT poses a challenging task and warrants further exploration.
- Compared to the SPT corpus, the existing downstream ABSA datasets cover fewer domains. This limits the in-depth analysis of aspect-level sentiments on different domains.

We believe that addressing the above limitations can facilitate the development of SPT.

## References


Hao Chen, Zepeng Zhai, Fangxiang Feng, Ruifan Li, and Xiaojie Wang. 2022. Enhanced multi-channel graph convolutional network for aspect sentiment triplet extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2974–2985, Dublin, Ireland. Association for Computational Linguistics.

Shaowei Chen, Yu Wang, Jie Liu, and Yuelin Wang. 2021. Bidirectional machine reading comprehension for aspect sentiment triplet extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12666–12674.

Zhuang Chen and Tieyun Qian. 2020. Relation-aware collaborative learning for unified aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3685–3694, Online. Association for Computational Linguistics.

Lingjia Deng and Janyce Wiebe. 2015. Mpqa 3.0: An entity/event-level sentiment corpus. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1323–1328.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Shuai Fan, Chen Lin, Haonan Li, Zhenghao Lin, Jinsong Su, Hang Zhang, Yeyun Gong, Jian Guo, and Nan Duan. 2022. Sentiment-aware word and sentence level pre-training for sentiment analysis. *arXiv preprint arXiv:2210.09803*.

Zhifang Fan, Zhen Wu, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. 2019. Target-oriented opinion words extraction with target-fused neural sequence labeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2509–2518, Minneapolis, Minnesota. Association for Computational Linguistics.

Chenggong Gong, Jianfei Yu, and Rui Xia. 2020. Unified feature and instance based domain adaptation for aspect-based sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7035–7045, Online. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2019. An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 504–515, Florence, Italy. Association for Computational Linguistics.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, New York, NY, USA. Association for Computing Machinery.

Binxuan Huang and Kathleen M Carley. 2019. Syntax-aware aspect level sentiment classification with graph attention networks. 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 5469–5477.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. 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. Association for Computational Linguistics.

Pei Ke, Haozhe Ji, Siyang Liu, Xiaoyan Zhu, and Minlie Huang. 2020. SentiLARE: Sentiment-aware language representation learning with linguistic knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6975–6988, Online. Association for Computational Linguistics.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Xin Li and Wai Lam. 2017. Deep multi-task learning for aspect term extraction with memory interaction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2886–2892, Copenhagen, Denmark. Association for Computational Linguistics.

Zhengyan Li, Yicheng Zou, Chong Zhang, Qi Zhang, and Zhongyu Wei. 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*, pages 246–256, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bin Liang, Hang Su, Lin Gui, Erik Cambria, and Ruifeng Xu. 2022. Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. *Knowledge-Based Systems*, 235:107643.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. 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 188–197, Hong Kong, China. Association for Computational Linguistics.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8600–8607.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. 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. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Nasim Nouri, Franck Dernoncourt, Dejing Dou, and Thien Huu Nguyen. 2020. Introducing syntactic structures into target opinion word extraction with deep learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8947–8956, Online. Association for Computational Linguistics.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion Word Expansion and Target Extraction through Double Propagation. *Computational Linguistics*, 37(1):9–27.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. 2020. SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4067–4076, Online. Association for Computational Linguistics.

Peter Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang. 2020. Relational graph attention network for aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3229–3238, Online. Association for Computational Linguistics.

Qianlong Wang, Zhiyuan Wen, Qin Zhao, Min Yang, and Ruifeng Xu. 2021. Progressive self-training with discriminator for aspect term extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 257–268, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Zhen Wu, Chengcan Ying, Fei Zhao, Zhifang Fan, Xinyu Dai, and Rui Xia. 2020a. Grid tagging scheme for aspect-oriented fine-grained opinion extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2576–2585, Online. Association for Computational Linguistics.

Zhen Wu, Fei Zhao, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. 2020b. Latent opinions transfer network for target-oriented opinion words extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9298–9305.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019a. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2324–2335, Minneapolis, Minnesota. Association for Computational Linguistics.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019b. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2324–2335, Minneapolis, Minnesota. Association for Computational Linguistics.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2018. Double embeddings and CNN-based sequence labeling for aspect extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 592–598, Melbourne, Australia. Association for Computational Linguistics.

Lu Xu, Lidong Bing, Wei Lu, and Fei Huang. 2020a. Aspect sentiment classification with aspect-specific opinion spans. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3561–3567, Online. Association for Computational Linguistics.

Lu Xu, Yew Ken Chia, and Lidong Bing. 2021. Learning span-level interactions for aspect sentiment triplet extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4755–4766, Online. Association for Computational Linguistics.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020b. Position-aware tagging for aspect sentiment triplet extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2339–2349, Online. Association for Computational Linguistics.

Wei Xue and Tao Li. 2018. Aspect based sentiment analysis with gated convolutional networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2514–2523, Melbourne, Australia. Association for Computational Linguistics.

Yichun Yin, Furu Wei, Li Dong, Kaimeng Xu, Ming Zhang, and Ming Zhou. 2016. Unsupervised word and dependency path embeddings for aspect term extraction. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 2979–2985. AAAI Press.

Jianfei Yu, Chenggong Gong, and Rui Xia. 2021. Cross-domain review generation for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4767–4777, Online. Association for Computational Linguistics.

Ziqian Zeng, Wenxuan Zhou, Xin Liu, and Yangqiu Song. 2019. A variational approach to weakly supervised document-level multi-aspect sentiment classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 386–396, Minneapolis, Minnesota. Association for Computational Linguistics.

Chen Zhang, Qiuchi Li, and Dawei Song. 2019. Aspect-based sentiment classification with aspect-specific graph convolutional networks. 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 4568–4578, Hong Kong, China. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *arXiv preprint arXiv:2203.01054*.

Jie Zhou, Junfeng Tian, Rui Wang, Yuanbin Wu, Wenming Xiao, and Liang He. 2020. SentiX: A sentiment-aware pre-trained model for cross-domain sentiment analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 568–579, Barcelona, Spain (Online). International Committee on Computational Linguistics.

## Appendix for "An Empirical Study of Sentiment-Enhanced Pre-Training for Aspect-Based Sentiment Analysis"

We organize the appendix into six sections:

- Additional data statistics for our SPT corpus and the downstream ABSA tasks are presented in Appendix A;
- Supplementary description of our sentiment knowledge mining method is presented in Appendix B;
- Detailed implementation of our masking context is presented in Appendix C;
- Addition notes for existing SPT approaches are presented in Appendix D;
- Detailed hyper-parameters for SPT are presented in Appendix E; and
- Additional experimental results are presented in Appendix F.

## A Additional Data Statistics

### A.1 Statistic of the SPT Corpus

We collect user-generated reviews from Yelp and Amazon to build the SPT corpus. We record the reviews from Yelp as the `Restaurant` domain and the reviews from *Cell Phones and Accessories* of Amazon as the `Laptop` domain. These two domains contain 11 million reviews in total. The detailed statistic is presented in Table 9.

**Note for `Laptop` reviews**. In the Amazon reviews dataset, Laptop is a subcategory under Electronics. However, this Laptop subcategory only contains 220k reviews, significantly less than the 4 million reviews on Yelp (`Restaurant`). To address this issue, some researchers have included reviews from other similar domains. For instance, Yin et al. (2016) added Cell Phone reviews to the Laptop corpus, while Wang et al. (2021) used Cell Phone reviews as the unlabeled data for semeval-14lap. Cell Phone category contains 6 million reviews, which is large enough. Although the use of Cell Phone reviews as a substitute for Laptop reviews is not ideal, our analysis shows that the two domains share high similarity in aspect-level sentiment expression.

| Domain | Number |
|---|---|
| Restaurant | 4,870,209 |
| Laptop | 6,330,313 |
| Beauty | 235,188 |
| Fashion | 550,447 |
| Appliances | 343,902 |
| Arts-Crafts | 1,735,948 |
| Automotive | 4,705,809 |
| Books | 31,442,868 |
| CDs | 2,395,159 |
| Clothing | 20,510,586 |
| Music | 921,391 |
| Electronics | 12,878,998 |
| Gift-Cards | 77,502 |
| Grocery-Food | 3,195,107 |
| Kitchen | 14,278,474 |
| Industrial | 1,028,681 |
| Kindle-Store | 3,838,324 |
| Luxury-Beauty | 371,295 |
| Magazine | 57,940 |
| Movies | 4,679,076 |
| Musical-Instruments | 930,999 |
| Office-Products | 3,424,566 |
| Garden | 3,247,801 |
| Pet-Supplies | 4,386,656 |
| Prime-Pantry | 251,476 |
| Sports | 8,180,487 |
| Toys | 5,161,466 |
| Video-Games | 1,442,524 |
| Total | 141,473,192 |

Table 9: Statistic of the SPT Corpus.

### A.2 Statistic of the Downstream ABSA tasks

**Dataset for the AE and ASC tasks** is provided by Wang et al. (2017). We download it from https://github.com/yhcc/BARTABSA/tree/main/data/wang. Since there is no explicit validation set, we randomly split 20% of its training data as the validation set. Its statistic is presented in Table 10. Following Xu et al. (2019b), for the ASC task, we only train and test the model on those aspect terms with polarity `POS`, `NEU`, and `NEG`.

**Dataset for the AOE task** is provided by Fan et al. (2019). We download it from https://github.com/NJUNLP/TOWE. Following Fan et al. (2019), we randomly split 20% of its training set as the validation set. Its statistic is presented in Table 10.

**Dataset for the ASTE task** is ASTE-Data-v2, which is provided by Xu et al. (2020b). We derive

| Provider | Domain | Split | #R | #A | #O | #P/#T |
|---|---|---|---|---|---|---|
| Wang et al. (2017) | **Rest 14** | Train | 2436 | 2985 | 2772 | - |
| | | Dev | 608 | 714 | 712 | - |
| | | Test | 800 | 1134 | 1008 | - |
| | **Lap 14** | Train | 2439 | 1915 | 2003 | - |
| | | Dev | 609 | 458 | 501 | - |
| | | Test | 800 | 654 | 674 | - |
| Fan et al. (2019) | **Rest 14** | Train | 1300 | 2109 | 2165 | 2443 |
| | | Dev | 325 | 530 | 557 | 619 |
| | | Test | 500 | 864 | 888 | 1030 |
| | **Lap 14** | Train | 920 | 1308 | 1293 | 1495 |
| | | Dev | 231 | 318 | 332 | 376 |
| | | Test | 343 | 481 | 498 | 565 |
| Xu et al. (2020b) | **Rest 14** | Train | 1266 | 2051 | 2061 | 2338 |
| | | Dev | 310 | 500 | 497 | 577 |
| | | Test | 492 | 848 | 844 | 994 |
| | **Lap 14** | Train | 906 | 1280 | 1254 | 1460 |
| | | Dev | 219 | 295 | 302 | 346 |
| | | Test | 328 | 463 | 466 | 543 |
| | **Rest 15** | Train | 605 | 862 | 935 | 1013 |
| | | Dev | 148 | 213 | 236 | 249 |
| | | Test | 322 | 432 | 460 | 485 |
| | **Rest 16** | Train | 857 | 1198 | 1300 | 1394 |
| | | Dev | 210 | 296 | 319 | 339 |
| | | Test | 326 | 452 | 474 | 514 |

Table 10: Statistics of three ABSA datasets (Wang et al., 2017; Fan et al., 2019; Xu et al., 2020b). #R, #A, #O, #P, and #T represent the number of reviews, aspect terms, opinion terms, aspect-opinion pairs, and aspect sentiment triplets, respectively.

it from https://github.com/xuuuluuu/Position-Aware-Tagging-for-ASTE. Its statistic is presented in Table 10.

**Dataset for cross-domain ABSA** is provided by Gong et al. (2020). We derive it from https://github.com/NUSTM/BERT-UDA. Its statistic is presented in Table 11.

**MAMS Dataset** is presented by Jiang et al. (2019). We download it from https://github.com/siat-nlp/MAMS-for-ABSA. Its statistic is presented in Table 11.

## B Supplementary Description of Our Sentiment Knowledge Mining Method

### B.1 Mining Aspect & Sentiment Words

Existing SPT works have developed several knowledge-mining methods (Tian et al., 2020; Zhou et al., 2020; Ke et al., 2020; Li et al., 2021). They leverage part-of-speech patterns or sentiment lexicons to annotate sentiment words. Then, they

Cross-domain ABSA (Gong et al., 2020)

| Domain | Split | #Review | #Aspect |
|---|---|---|---|
| **Restaurant** | Train | 3877 | 3626 |
| | Test | 2158 | 1994 |
| **Laptop** | Train | 3045 | 1845 |
| | Test | 800 | 467 |
| **Device** | Train | 2557 | 1394 |
| | Test | 1279 | 691 |
| **Service** | Train | 1492 | 1729 |
| | Test | 747 | 825 |

MAMS (Jiang et al., 2019)

| | Split | #Review | #Aspect |
|---|---|---|---|
| **ATSA** | Train | 4297 | 11182 |
| | Dev | 500 | 1332 |
| | Test | 500 | 1336 |
| **ACSA** | Train | 3149 | 7090 |
| | Dev | 400 | 888 |
| | Test | 400 | 901 |

Table 11: Statistics of the cross-domain ABSA dataset (Gong et al., 2020) and MAMS (Jiang et al., 2019) dataset.

treat the nearest nouns to the sentiment words as aspect words or build an aspect lexicon based on the aspect annotations of the existing downstream datasets. However, these knowledge-mining methods either lack domain adaptability or are unscalable. Therefore, this paper develops an effective knowledge-mining method.

**Hypothesis.** Any two words in the same sentence are connected by a syntactic path. This paper denotes a syntactic path as a sequence of dependency relations and part-of-speech tags. For example, given the sentence "*we had a lamb pie pizza that was awful*", the syntactic path from *pizza* to *awful* is denoted as $(\texttt{NOUN}, \xrightarrow{relcl}, \texttt{VERB}, \xrightarrow{acomp}, \texttt{ADJ})$. Qiu et al. (2011) observe that there are some syntactic paths linking aspect words and sentiment words. Based on this observation, we assume that there exist lexicons $\mathcal{L}_{\mathcal{A}}, \mathcal{L}_{\mathcal{S}}$ and aspect-sentiment path set $\mathcal{P}_{\mathcal{AS}}$ that satisfy:

If $w_i$ and $w_j$ are linked by path $p$, then

$$w_i \in \mathcal{L}_{\mathcal{A}} \text{ and } p \in \mathcal{P}_{\mathcal{AS}} \Longrightarrow w_j \in \mathcal{L}_{\mathcal{S}}, \quad (18)$$

$$w_i \in \mathcal{L}_{\mathcal{A}} \text{ and } w_j \in \mathcal{L}_{\mathcal{S}} \Longrightarrow p \in \mathcal{P}_{\mathcal{AS}}, \quad (19)$$

We leverage this assumption to mine lexicons and path sets.

**Initialization.** We initialize the sentiment lexicon

with adjectives of MPQA[6] and initialize the aspect lexicon as an empty set. We empirically initialize four path sets with several simple paths:

$$\mathcal{P}^{(0)}_{\mathcal{AA}} = \{(\texttt{NOUN}, \xrightarrow{conj}, \texttt{NOUN}), (\texttt{NOUN}, \xleftarrow{conj}, \texttt{NOUN})\},$$
$$\mathcal{P}^{(0)}_{\mathcal{SS}} = \{(\texttt{ADJ}, \xrightarrow{conj}, \texttt{ADJ}), (\texttt{ADJ}, \xleftarrow{conj}, \texttt{ADJ})\},$$
$$\mathcal{P}^{(0)}_{\mathcal{AS}} = \{(\texttt{NOUN}, \xleftarrow{nsubj}, \texttt{AUX}, \xrightarrow{acomp}, \texttt{ADJ})\},$$
$$\mathcal{P}^{(0)}_{\mathcal{SA}} = \{(\texttt{ADJ}, \xleftarrow{acomp}, \texttt{AUX}, \xrightarrow{nsubj}, \texttt{NOUN})\}.$$

**Expectation.** We first leverage the two lexicons to annotate aspect words and sentiment words in the sentences. Then, we treat these annotations as conditions to further annotate aspect and sentiment words in the sentences using the four path sets through Equation (18). We merge the two annotations as the final annotation of this step.

**Maximization.** For each word $w_i$, we count its occurrences of being annotated as the aspect word in the E-step, denoted as $\#w_i\_asp$. We also count its occurrences in the corpus, denoted as $\#w_i$. Thus, we calculate the aspect score of a word as follows:

$$aspect\text{-}score(w_i) = \frac{\#w_i\_asp}{\#w_i}. \quad (20)$$

We add those words whose aspect scores are greater than the given threshold $\alpha_{\mathcal{A}}$ to the aspect lexicon $\mathcal{L}_{\mathcal{A}}$. The expansion of the sentiment lexicon is similar.

For each path $p$, we count the annotations of all the word pairs $(w_i, w_j)$ linked by this path. The aspect-sentiment score of the path is calculated by:

$$as\text{-}score(p) = \frac{\#w_i\_asp * \#w_j\_sen}{\#w_i\_asp}, \quad (21)$$

where $\#w_j\_sen$ denotes its occurrences of being annotated as the sentiment word. We add those paths whose aspect-sentiment score is greater than the given threshold $\alpha_{\mathcal{AS}}$ to $\mathcal{P}_{\mathcal{AS}}$. The other three path sets are expanded similarly.

For each domain, we run the EM algorithm on 500k reviews to learn the parameters and then annotate all reviews.

[6] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

### B.2 Polarity Assignment

We leverage the reviews' rating scores for polarity assignments. We empirically treat 5-star-rated reviews as positive reviews and (1,2,3)-star-rated as negative reviews. For a sentiment word $w_i$, we count its occurrences in positive and negative reviews, denoted as $\#w_i\_pos$ and $\#w_i\_neg$. Thus the polarity score of this word can be calculated by:

$$pmi\_so(w_i) = \log \frac{\#w_i\_pos * \#neg}{\#w_i\_neg * \#pos}, \quad (22)$$

where $\#pos$ and $\#neg$ denote the total number of positive and negative reviews. Equation (22) is derived from Pointwise Mutual Information (PMI) (Turney, 2002).

For each domain, we calculate the polarity score of each sentiment word on all reviews. Then, we empirically assign those sentiment words whose polarity scores are greater than 0.2 as `POS` and those sentiment words whose polarity scores are less than -0.2 as `NEG`.

## C Detailed Implementation of Masking Context

We leverage aspect words to locate the sentiment-dense segments of the review and improve the masking probability of their contexts.

**Masking Context by Normal Distribution.** Suppose there is only one aspect in the review, and its position is $t$. We use the normal distribution $\mathcal{N}(t, \sigma)$ for the masking probability assignment, where $\sigma$ is a hyper-parameter. This could be formulated by:

$$f_{norm}(i; t, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(i-t)^2}{2\sigma^2}\right).$$
$$p_{norm}(i) = \begin{cases} f_{norm}(i; t, \sigma) & i \neq t, \\ 0 & i = t. \end{cases}$$

Actually, a review often contains more than one aspect word. Therefore, we sample $k$ aspect words for each review, repeat the probability calculation $k$ times, and pick up the maximum probability for each word. $k$ is related to the length of the review:

$$k = Tz, \quad (23)$$

where $T$ is the length, and $z$ is a hyper-parameter. Finally, we perform normalization to ensure that the masked part is 15% of the review.

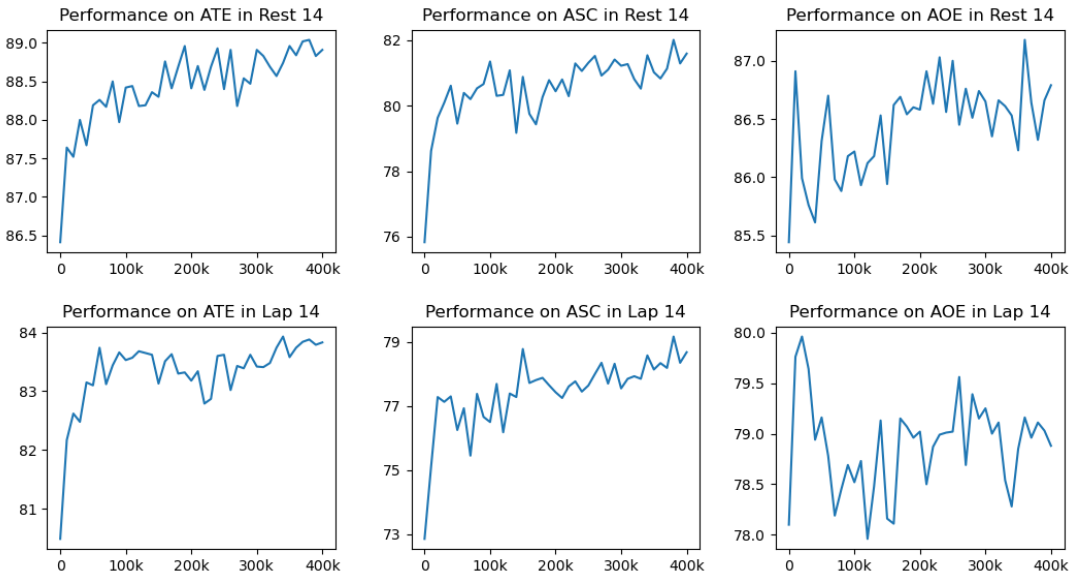


Figure 4: Performance of three downstream ABSA tasks with different pre-training steps.

| Hyper-parameter | SPT (10k) | SPT (400k) |
|---|---|---|
| Pre-training Data | `Restaurant` and `Laptop` | All 28 domains |
| Batch Size | 1k | 1k |
| Learning Rate | 2e-4 | 1e-4 |
| Max Step | 10k | 400k |
| Learning Rate Decay | Linear | Linear |
| Weight Decay | 0.01 | 0.01 |
| Adam $\epsilon$ | 1e-6 | 1e-6 |
| Adam $\beta_1$ | 0.9 | 0.9 |
| Adam $\beta_2$ | 0.98 | 0.98 |
| Gradient Clipping | 1.0 | 1.0 |

Table 12: Hyper-parameters for pre-training.

**Masking Context by Geometric Distribution.** We also use geometric distribution for the masking probability assignment:

$$f_{geo}(i;p) = (1-p)^{i-1}p. \tag{24}$$

$$p_{geo}(i) = \begin{cases} f_{geo}(t-i;p) & i < t, \\ 0 & i = t, \\ f_{geo}(i-t;p) & i > t. \end{cases} \tag{25}$$

where $p$ is a hyper-parameter. We also sample $k$ aspect words and perform normalization.

We find that both masking context methods are highly sensitive to hyper-parameters. In our experiment, we set $\sigma = 6$, $p = 0.4$, and $z = 0.1$.

## D Additional Notes for Existing SPT Approaches

**Note for SEKP.** Tian et al. (2020) propose aspect-sentiment pair prediction to capture the dependency between aspect and sentiment. They regard aspect-sentiment pair prediction as a multi-label classification task and further transform it into multiple binary classification tasks. However, their implementation only includes positive samples but ignores negative samples when calculating the loss. Therefore, we adopt a different implementation to correct this mistake. This implementation is described in Equation 12 and 13. Besides, since our pre-training corpus contains aspect annotation, we regard a sentiment word with its nearest aspect word as an aspect-sentiment pair.

**Note for Pre-training Step.** Existing SPT works have different setups. We estimate their computation based on batch size, maximum text length, model architecture, and training step. Then, we calculate the training step to reach this computation under our settings.

## E Hyper-Parameters

We list the detailed hyper-parameters of SPT in Table 12. In our SPT, we integrate ASPECTWORD, REVIEWRATING, and SYNTAX in pre-training. Therefore, the loss for SPT is calculated by:

$$\mathcal{L}_{SPT} = \mathcal{L}_{MLM} + \alpha_1 \mathcal{L}_{RAT} + \alpha_2 \mathcal{L}_{SYN}, \tag{26}$$

$$\mathcal{L}_{SYN} = \alpha_3 \mathcal{L}_{PoS} + \alpha_4 \mathcal{L}_{DIR} + \alpha_5 \mathcal{L}_{DIS}. \tag{27}$$

| Backbone | Source → Target Pairs | | | | | | | | | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | D→R | D→S | L→R | L→S | R→D | R→L | R→S | S→D | S→L | S→R | |
| BERT | 47.62 | 36.16 | 36.72 | 30.53 | 38.53 | 38.29 | 26.10 | 39.30 | 39.81 | 56.82 | 38.99 |
| +SKEP | 53.08 | 38.12 | 51.80 | 28.19 | 44.09 | 47.28 | 25.92 | 39.52 | 41.25 | 54.48 | 42.37 |
| +SENTILARE | 51.44 | 33.49 | 48.52 | 30.78 | 43.07 | 47.19 | 28.70 | 40.13 | 41.91 | 56.64 | 42.19 |
| +SCAPT$_{\text{LAP}}$ | 47.46 | 36.29 | 46.07 | 29.74 | 39.75 | 48.39 | 32.22 | 40.58 | 42.61 | 58.99 | 42.21 |
| +SCAPT$_{\text{REST}}$ | 54.22 | 36.76 | 40.55 | 31.53 | 40.61 | 40.14 | 29.02 | 39.74 | 39.22 | 58.22 | 41.07 |
| +SENTIX | 52.89 | 39.26 | 45.65 | 34.21 | 41.77 | 44.08 | 29.16 | 40.72 | 41.06 | 58.19 | 42.70 |
| +BERT$_{\text{REVIEW}}$ | 54.99 | 42.77 | 51.29 | 33.17 | 45.49 | 50.08 | 32.82 | 41.48 | 42.41 | 61.17 | 45.57 |
| **+Our SPT** | 55.02 | 36.52 | 52.06 | 30.25 | 44.05 | 47.81 | 27.63 | 41.20 | 40.40 | 55.96 | 43.09 |
| **+Our SPT** | 57.30 | 36.14 | 56.45 | 33.22 | 48.10 | 53.17 | 31.95 | 39.90 | 40.56 | 59.33 | **45.61** |

Table 13: Full results for cross-domain Aspect term Extraction ($F_1$-score, %). These results are the average of 10 runs.

| Backbone | Source → Target Pairs | | | | | | | | | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | D→R | D→S | L→R | L→S | R→D | R→L | R→S | S→D | S→L | S→R | |
| BERT | 42.57 | 33.29 | 35.94 | 29.48 | 35.25 | 37.06 | 25.35 | 35.59 | 36.64 | 50.08 | 36.13 |
| +SKEP | 49.36 | 32.87 | 49.44 | 28.57 | 43.45 | 46.01 | 25.76 | 37.51 | 41.13 | 50.29 | 40.44 |
| +SENTILARE | 49.84 | 32.86 | 49.34 | 29.10 | 43.54 | 47.21 | 27.31 | 38.87 | 41.37 | 52.74 | 41.22 |
| +SCAPT$_{\text{LAP}}$ | 44.94 | 33.54 | 44.31 | 29.14 | 37.30 | 45.98 | 30.28 | 38.40 | 40.93 | 54.16 | 39.90 |
| +SCAPT$_{\text{REST}}$ | 52.58 | 34.52 | 40.72 | 30.42 | 38.38 | 37.90 | 26.45 | 37.72 | 37.86 | 54.52 | 39.11 |
| +SENTIX | 50.09 | 34.27 | 45.26 | 31.74 | 40.22 | 45.60 | 27.43 | 38.84 | 39.88 | 54.31 | 40.76 |
| +BERT$_{\text{REVIEW}}$ | 53.20 | 38.69 | 48.82 | 33.58 | 44.37 | 49.17 | 33.64 | 38.66 | 40.60 | 56.98 | 43.77 |
| **+Our SPT** | 52.64 | 34.40 | 51.44 | 29.28 | 43.56 | 46.66 | 26.05 | 39.25 | 40.19 | 52.40 | 41.59 |
| **+Our SPT** | 55.70 | 33.81 | 55.90 | 31.50 | 46.95 | 52.33 | 30.13 | 38.71 | 41.01 | 55.51 | **44.16** |

Table 14: Full results for cross-domain end-to-end Aspect-Based Sentiment Analysis ($F_1$-score, %). These results are the average of 10 runs.

We set $\alpha_1 = 1.2$, $\alpha_2 = 1.2$, $\alpha_3 = 0.8$, $\alpha_4 = 1.0$, $\alpha_5 = 0.3$.

## F Additional Results

**Effect of Pre-training Step**. The pre-training step is the most critical hyper-parameter in SPT. We show its effect on the downstream ABSA tasks in Figure 4. We can see that the performance on AE and ASC generally increases with the number of pre-training steps. Although the increase is slowly decreasing, it can be expected that the performance will continue to increase if pre-training continues on the basis of SPT (400k). However, this trend is not obvious on AOE. In particular, the performance in `Laptop-14` degrades as the pre-training step increases. We speculate that this is due to the differences between domains, which are more pronounced for the AOE task. Not only does the wording of the opinion term differ across domains, but the dependency between the aspect term and opinion term also differs across domains. These differences deserve further exploration and analysis in future works.

**Full Results on Cross-domain ABSA**. We present the full results on two cross-domain ABSA tasks in Table 13 and 14.

## ACL 2023 Responsible NLP Checklist

### A For every submission:

☑ A1. Did you describe the limitations of your work?
*Section of Limitations*

☐ A2. Did you discuss any potential risks of your work?
*Not applicable. Left blank.*

☑ A3. Do the abstract and introduction summarize the paper's main claims?
*Section 1*

☑ A4. Have you used AI writing assistants when working on this paper?
*1. Grammarly. We use it to correct grammatical errors. 2. DeepL. We use it for translation between languages.*

### B ☑ Did you use or create scientific artifacts?

*Appendix A.2.*

☑ B1. Did you cite the creators of artifacts you used?
*Appendix A.2.*

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
*The artifacts we use have no terms attached to them but can be used in various scientific studies by default when provided they are cited.*

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
*The artifacts we use do not specifically state their intended use*

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
*Not applicable. Left blank.*

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*Not applicable. Left blank.*

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
*Appendix A.2.*

### C ☑ Did you run computational experiments?

*Section 3*

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
*Appendix E.*

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*Appendix E.*

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
*Section 3*

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
*Section 2.1.2*

**D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*Left blank.*

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
*Not applicable. Left blank.*

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
*Not applicable. Left blank.*

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
*Not applicable. Left blank.*

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*Not applicable. Left blank.*

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*Not applicable. Left blank.*