Modeling Span-Level Interactions with Boundary Probabilities for Targeted Sentiment Analysis

Anonymous ACL submission

Abstract

Aspect extraction and sentiment prediction are two important tasks for targeted sentiment analysis. Recently, some span-based methods have gained great attention to capture the associations between the two tasks in this domain, where they first extract aspects by detecting aspect boundaries and then predict the span-level sentiments. Most existing studies on modeling the inter-task interactions either share the input representation of the two tasks at the encoding layer, or approximate the output representation of the two tasks at the task layer. Both of them focus on modeling sentence-level correlations between these two tasks, leading to insufficient inter-task feature interactions. Since the aspect-level features are also crucial to connect these two tasks, thus, different from previous approaches, in this paper, we propose to model the span-level interactions with boundary probabilities (SIBP) to explicitly consider the inner correlations for these two tasks. Specifically, we use the predicted boundary probabilities of aspects to generate all possible spans as input of the sentiment prediction module, such that the sentiment information can be backpropagated into the boundary detection process in a fully differentiable manner. Further, we devise an alternate learning strategy to take the best of both tasks between predicted aspects and real aspects. This strategy not only guides sentiment prediction more properly but also improves computational efficiency. Moreover, to predict the boundary probabilities of the aspects more accurately, we design a semantic compatibility mechanism. Finally, we conduct extensive experiments on three real-world datasets to demonstrate the model's superiority.

1 Introduction

Targeted sentiment analysis (Pang et al., 2008; Bing, 2012) has long been an important natural language processing task, which involves of two sub-problems: (1) **aspect extraction** (Jakob and Gurevych, 2010; Poria et al., 2016; Karimi et al., 2021), which aims to identify the aspects in a sentence, and (2) **sentiment prediction** (Jiang et al., 2011; Lin et al., 2019; Karimi et al., 2021), which predicts the sentiment polarity of a given aspect (e.g., positive, negative, and neutral).

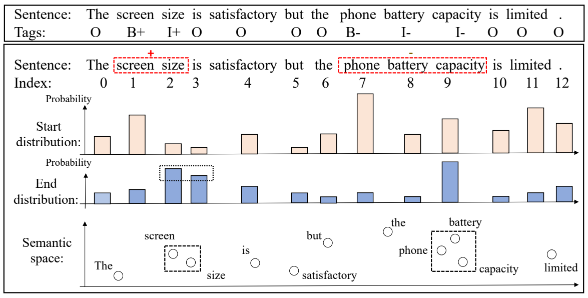

Sentence: The screen size is satisfactory but the phone battery capacity is limited .
Tags: O B+ I+ O O O O B- I- I- O O O
+
-
Sentence: The screen size is satisfactory but the phone battery capacity is limited .
Index: 0 1 2 3 4 5 6 7 8 9 10 11 12
Probability
Start distribution:
Probability
End distribution:
Semantic space:
The
screen
size
is
but
satisfactory
the
battery
phone
capacity
limited

Figure 1: The upper part shows an example of tagging-based models. The bottom part illustrates the span-based methods, where we present an example semantic space for different words.

Generally, previous research in this domain can be largely classified into two types: tagging-based and span-based models. Tagging-based models (Luo et al., 2019; Li et al., 2019; Wang et al., 2021) combine two different problems via designing the combination tags. For example, in the sentence of Figure 1, each tag is composed of two parts, the first part is used to label the aspect words, while the second part indicates the word sentiment, e.g., "B-" means the beginning of a negative aspect. Despite effectiveness, tagging-based methods need to search huge word index spaces and cannot guarantee sentiment consistency (Hu et al., 2019). To alleviate these shortcomings, the span-based models (Hu et al., 2019; Lin and Yang, 2020) are proposed, where the aspect is detected by directly predicting the boundary distributions, and the sentiment is classified based on all the words between the boundaries. Instead of labeling each word sentiment sequentially, span-based models predict a unified sentiment for all the words in the aspect. In this work, we focus on the span-based models.

Recently, some span-based approaches (Lv et al., 2021; Chen et al., 2022) start considering the interactions between these two tasks to achieve strong performance due to their inner correlation. Most prior works either focus on input-side interactions, sharing partial sentence encodings, or they consider the output-side interactions, forcing the output representations of the two tasks to be close. Unfortunately, both of them model the inter-task interactions relying on sentence-level features, ignoring aspect-level feature interactions. Indeed, the aspect-level features play an essential role when connecting two tasks. On the one hand, sentimental expressions can be understood precisely if the desired aspects are specific, e.g., “my new house is close to **sports park**." and “my new house is close to **stinky ditch**.", the sentimental words “close to" show a positive attitude towards “sports park” but the negative attitude towards “stinky ditch”. On the other hand, aspects can also be quickly detected if we know where the sentimental expressions locate, e.g., if a negative sentimental words “acid” appears in a hotel review, there is likely an aspect word related to “**food**". Previous span-based models usually neglect such sequential connections between the two tasks, which may limit their performances.

Inspired by the above analysis, in this paper, we propose SIBP, to model span-level interactions with boundary probabilities for solving the task of targeted sentiment analysis. Specifically, we utilize the boundary distribution predicted from the aspect extraction model to generate all possible spans as input for sentiment prediction. Based on this strategy, the aspect information can guide the sentiment prediction task, and the sentiment information can provide helpful signals via backpropagation for the aspect extraction task in a fully differentiable manner. Then, we devise an alternate learning strategy to achieve a trade-off between predicted aspects and real aspects. This strategy not only ensures that both tasks are well learned but also reduces the computational overhead.

Moreover, we also design an advanced semantic compatibility mechanism to combine the position and semantic information for detecting aspects in a unified manner. Most previous models predict the start and end boundaries independently. And to the best of our knowledge, there is no specific mechanism to guarantee that the words between the boundaries can form an aspect, which may negatively impact the capability of aspect extraction. For example, in Figure 1, the end boundary distribution implies that the 2nd and 3rd words have similar chances to become the end of an aspect. However, from the semantic perspective, the word “is” can be irrelevant to the former words, and less likely to be the end of an aspect. Semantic information is useful in checking whether different words can form a “semantically reasonable” aspect, which may play a complementary role together with the boundary distributions. Finally, extensive experiments on three datasets and several baseline models prove the effectiveness of our proposed framework. The major contributions of this paper are summarized as follows:

- We propose a span-level interaction method with boundary probabilities (SIBP) for solving the task of targeted sentiment analysis.
- To predict the boundary probabilities of aspects more accurately, we design a semantic compatibility mechanism, which provides complementary signals to boundary detection.
- We validate the effectiveness of SIBP on three real-world datasets, and results indicate our model outperforms a variety of state-of-the-art baselines.

2 Related Work

The aspect extraction and sentiment prediction are the utmost trending tasks of targeted sentiment analysis. Recently, considering the fact that both aspect and sentiment can provide strong mutual indications for each other, research studies have been performed to devise tagging-based and span-based solutions for extracting aspect-sentiment jointly. Here, we briefly review these models in the following.

Tagging-based methods. Many researchers adopted the tagging-based strategy to solve these two tasks jointly, which basically regard the problem as a sequential labeling process (Mitchell et al., 2013). The aspect and sentiment information is scattered onto each word, and the sentence is labeled with tags. With this formulation, early models (Zhang et al., 2015; Li and Lu, 2017) leverage Conditional Random Field (CRF) (Lafferty et al., 2001) to autoregressively predict the tags based on handcrafted linguistic features. To exploit the powerful representations of encoders (e.g., LSTM, CNN, BERT), recent studies (Li et al., 2019; Mao

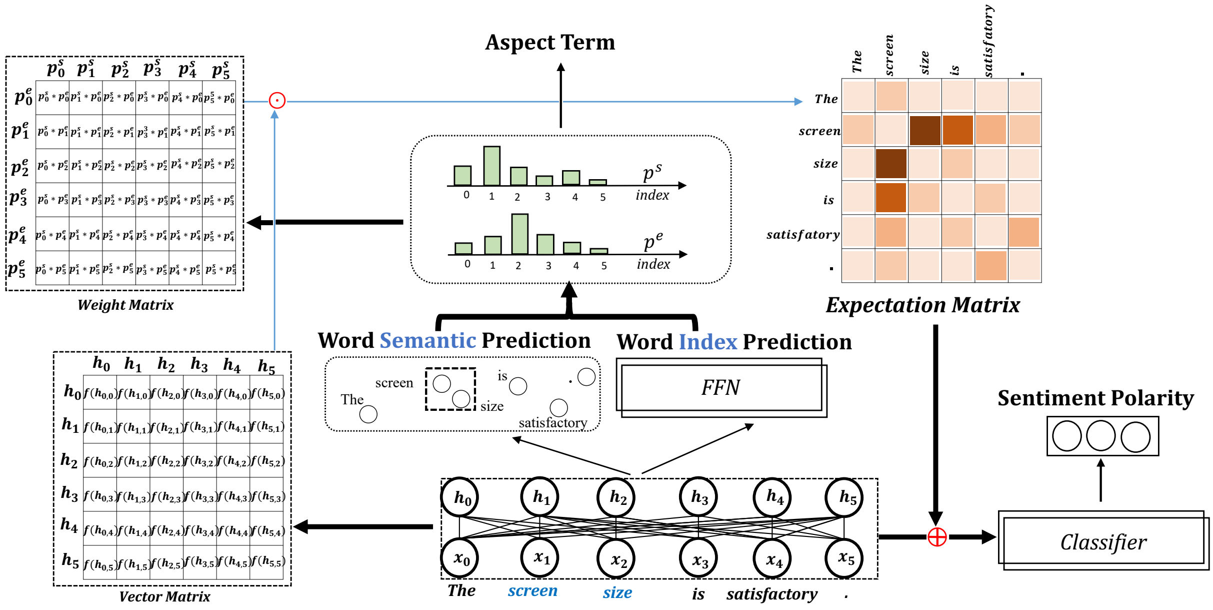


Figure 2: A depiction of the SIBP model. The model is specifically designed to extract aspects and sentiments from reviews by performing the following operations: (1) Encoding the input sentence representations for aspect extraction and sentiment prediction through the BERT encoder; (2) Generating the boundary probabilities (i.e., start and end) through word semantic prediction module and word index prediction module; (3) Generating expectation matrix though weight matrix and vector matrix; and (4) Incorporating the expectation features of aspect to sentiment predicting. Precisely, in the expectation matrix, the darker the color, the greater the weight on the span.

et al., 2021; Zhang et al., 2022) have paid more attention to designing neural networks for predicting the tags sequentially. Further, several models (Hu et al., 2019; Chen et al., 2020; Yan et al., 2021) leverage advanced learning technology (e,g., graph representation learning and multi-task learning) to extract aspects and predict sentiments jointly. Although the above models have achieved many successes, the main problem is that they disperse the whole aspect sentiment to each word within it, and basically relax the constraint of sentiment consistency, which increases the sentiment ambiguity.

Span-based methods. Meanwhile, some of the researchers implemented the span-based strategies (Hu et al., 2019; Zhou et al., 2019), that is, first extracting aspects by detecting aspect boundaries and then predicting the span-level sentiments. Hu et al. (2019) proposed an extract-then-classify framework that first extracts targets with a heuristic decoding algorithm, and then correspondingly classifies the sentiments. In this work, a shared input embedding is introduced to consider the correlations between two tasks. To extend the above work, Lin and Yang (2020) devised a shared-private representation method that implicitly incorporates the shared and private information in sentence embedding between two tasks to improve task results. Lv et al. (2021) explicitly constructed dual gated recurrent units and an interaction layer to model the inter-task connections. Further, Chen et al. (2022) developed an interactive network in a hierarchical manner (i.e., input-side interactions and output-side interactions) that takes explicit cooperation from aspect extraction and sentiment prediction to enhance the model performance.

Unlike existing works on aspect extraction and sentiment prediction that model the inter-task interactions relying on sentence-level features. In this work, we novelly propose to model the span-level interactions with boundary probabilities for these two tasks. And we devise an alternate learning strategy to achieve a trade-off between predicted aspects and real aspects. Moreover, this study also designs a semantic compatibility mechanism to extract the aspects more accurately. Compared with the current endeavors, our models yield better performance.

3 Methodology

The overall framework is depicted in Figure 2. We can see that the sentiment prediction module takes the predicted aspects as input, and outputs the sentiment of the aspects. Under the guidance of the sentiment information, the boundary distributions are regularized and the aspect is extracted in a more accurate manner.

Given a sentence $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ where x_i is the i^{th} token in $\boldsymbol{x}$, and n is the length of sentence, the main task is to extract all possible aspect terms (s, e) (the start and end indexes of aspect term), and predict corresponding sentiments (i.e., positive, neutral, and negative) in $\boldsymbol{x}$.

In order to represent the semantics more accurately, we use the well-known bidirectional encoder representation from transformers (BERT) (Devlin et al., 2019) to process the raw input. For a sentence $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$, we suppose the output from BERT is $\boldsymbol{H} = \{h_0, h_1, \ldots, h_n\} \in \mathbb{R}^{n \times d}$, where d is the embedding size, n is the sentence's length.

3.1 Aspect Boundary Prediction

In previous models (Hu et al., 2019; Lin and Yang, 2020), the aspects are detected by predicting the start and end boundaries through a linear transformation operation. To be specific, the distribution of the start boundary is predicted as,

$$\boldsymbol{g}_s = \boldsymbol{H}\boldsymbol{W}_1 + \boldsymbol{b}_1 \quad , \qquad \hat{\boldsymbol{p}}_s = \text{softmax}(\boldsymbol{g}_s), \quad (1)$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{d \times 1}$ and $\boldsymbol{b}_1 \in \mathbb{R}^n$ are the weighting and bias parameters, softmax$(\cdot)$ normalizes the input into a distribution. In this equation, the length of $\hat{\boldsymbol{p}}_s$ is the same as the sentence's, and each element in $\hat{\boldsymbol{p}}_s$ represents the probability of a word being the start of an aspect. The end distribution can be derived in a similar manner, and we define it as $\hat{\boldsymbol{p}}_e$. The learning objective for the boundary distributions is,

$$\mathcal{J}_{ae} = -\sum_{i=1}^{n} \{\boldsymbol{p}_{i,s}^T \log(\hat{\boldsymbol{p}}_{i,s}) + \boldsymbol{p}_{i,e}^T \log(\hat{\boldsymbol{p}}_{i,e})\}, \quad (2)$$

where $\boldsymbol{p}_{i,s}^T \in \mathbb{R}^n$ and $\boldsymbol{p}_{i,e}^T \in \mathbb{R}^n$ are the ground truths of the boundaries (i.e., 0-1 vectors), $\hat{\boldsymbol{p}}_{i,s}$ and $\hat{\boldsymbol{p}}_{i,e}$ are the predicted boundary distributions, n is the length of sentence.

3.2 Sentiment Prediction

For a given aspect, the sentiment is predicted based on the words between its boundaries. Suppose the start and end boundaries are s and e, respectively, we calculate a summarized vector α using the attention mechanism over tokens in its corresponding bound (s,e)[1], thus, the sentiment is predicted as,

$$\alpha = \text{softmax}\left(\boldsymbol{W}_\alpha \boldsymbol{H}_{s \to e}\right), \quad (3)$$

$$\begin{aligned} \hat{\boldsymbol{y}} =& \text{softmax}(\boldsymbol{W}_3 \tanh(\boldsymbol{W}_4 \cdot \mathcal{F}) \\ =& \text{softmax}(\boldsymbol{W}_3 \tanh(\boldsymbol{W}_4 \sum_{i=s}^{e} \alpha_i \boldsymbol{H}_i)), \end{aligned} \quad (4)$$

where $\mathcal{F}$ is the aspect vector, and $\boldsymbol{W}_\alpha$, $\boldsymbol{W}_3$ and $\boldsymbol{W}_4$ are weighting parameters. This sentiment prediction method is actually used by most of the previous models (Lin et al., 2019; Tian et al., 2021b). Then, the parameters are learned based on cross-entropy,

$$\mathcal{J}_{sp} = -\sum_{i=1}^{N} \boldsymbol{y}_i^T \log \hat{\boldsymbol{y}}_i \;= -\sum_{i=1}^{N}\sum_{j=1}^{K} y_{i,j} \log \hat{y}_{i,j}, \quad (5)$$

where N is the total number of training samples, K is the number of sentiment types, $\hat{\boldsymbol{y}}_i$ is the predicted sentiment distribution, and $\boldsymbol{y}_i^T$ is the corresponding ground truth.

3.3 Modeling Span-Level Interactions

Most existing approaches capture the inter-task correlations based on sentence-level features. They model these interactions from the input-side or output-side to share sentence representation, both of which fail to sufficiently model connections between aspect extraction and sentiment prediction.

As mentioned above, the aspect can provide useful signals for sentiment prediction. Conversely, sentimental features are also beneficial to the extraction of aspect boundaries. However, the sentiment is predicted based on the ground truth boundaries in previous works, which basically separates the sequential connections between these two tasks. Therefore, in this paper, we use the aspect features to supervise the sentiment predicting process. On the one hand, aspect extraction can sequentially deliver useful supervised signals for sentiment prediction. On the other hand, sentiment information can influence aspect detection through backpropagation. To this end, we do not use the boundary ground truth as the input, instead, we derive an expectation of the word embedding based on the boundary distributions, thus, the Equation 4 can be rewritten to,

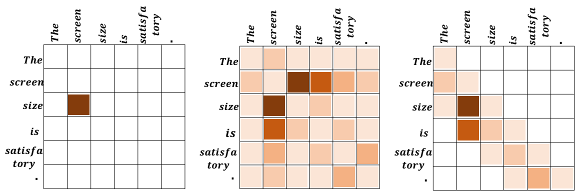
$$\begin{aligned} \hat{\boldsymbol{y}} &= \text{softmax}(\boldsymbol{W}_3 \tanh(\boldsymbol{W}_4 \cdot \hat{\mathcal{F}}) \\ &= \text{softmax}(\boldsymbol{W}_3 \tanh(\boldsymbol{W}_4 \mathbb{E}_{(i \sim \hat{\boldsymbol{p}}_s, j \sim \hat{\boldsymbol{p}}_e)}[f(\boldsymbol{H}_{i \to j})])), \end{aligned} \quad (6)$$

where $\hat{\mathcal{F}}$ is the expected aspect vector, and f is an aggregation function, and we specify it as,

$$f(\boldsymbol{H}_{i \to j}) = \sum_{k=i}^{j} \boldsymbol{H}_k, \quad (7)$$

Since $\hat{\boldsymbol{p}}_s$ and $\hat{\boldsymbol{p}}_e$ are both discrete distributions, we have the following derivation,

[1] Please note that (s, e) is ground truth boundaries.



(a) $\mathcal{F}$ (b) $\hat{\mathcal{F}}$ (c) $\hat{\mathcal{F}}_o$

Figure 3: The matrix representation about the sentence "The screen size is satisfactory.". The matrix on the left represents using the real aspect as input, the middle matrix denotes using the boundary probabilities expectation of the generated aspect as input, and the right means using the optimized aspect expectation as input.

$$\begin{aligned}\hat{\mathcal{F}} &= \mathbb{E}_{(i\sim\hat{\boldsymbol{p}}_s, j\sim\hat{\boldsymbol{p}}_e)}[f(\boldsymbol{H}_{i\rightarrow j})] \\ &= \mathbb{E}_{i\sim\hat{\boldsymbol{p}}_s}[\sum_{j=1}^{n}\hat{p}_{e,j} f(\boldsymbol{H}_{i\rightarrow j})] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\hat{p}_{s,i}\ \hat{p}_{e,j} f(\boldsymbol{H}_{i\rightarrow j}) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\hat{p}_{s,i}\ \hat{p}_{e,j}\sum_{k=i}^{j}\boldsymbol{H}_k,\end{aligned} \tag{8}$$

where we assume n is the sentence's length.

In the training process, in order to more accurately model the expected output of each aspect, we split the sentences containing multiple aspects into multiple sentences (i.e., each sentence contains only one aspect), which can ensure these aspect words in the same sentence are not disturbed during learning. However, during inference, we can still extract multiple aspects and corresponding sentiments simultaneously if the sentence contains multiple aspects.

Appendix A proves that the optimization objective does not change when changing the partition of the training set.

Remark.

(1) Comparison with $\mathcal{F}$. For a more intuitive understanding of leveraging boundary probabilities as input, we compare it with the previous approach of leveraging real aspect as input in Figure 3a and 3b (i.e., $\mathcal{F}$ and $\hat{\mathcal{F}}$). More specifically, $\mathcal{F}$ can be regarded as a 0-1 matrix, and the real aspect is used as input to supervise the learning of the sentiment predicting, which basically cuts off the connection between the two tasks. $\hat{\mathcal{F}}$ is a differentiable probability matrix, the sentiment information can be backpropagated into the boundary detection process via $\hat{\boldsymbol{p}}_s$ and $\hat{\boldsymbol{p}}_e$ with our method, which captures the sequential nature between different tasks.

(2) Optimizing $\hat{\mathcal{F}} \rightarrow \hat{\mathcal{F}}_o$. In practice, during training, to compute Equation 8 (i.e., $\hat{\mathcal{F}}$), one needs to traverse the whole sentence, and the computational complexity is $\mathcal{O}(n^2)$, which is very difficult to calculate[2]. Fortunately, this complexity can be reduced based on two facts: (i) The end point is not smaller than the start point. (ii) The length of an aspect term is usually not large. Then we revise Equation 8 as $\hat{\mathcal{F}}_o$ (see Figure 3c), that is,

$$\hat{\mathcal{F}}_o = \sum_{i=1}^{n}\sum_{j=i}^{i+h}\hat{p}_{i,s}\hat{p}_{e,j}\sum_{k=i}^{j}\boldsymbol{H}_k, \tag{9}$$

where h is the maximum length of the aspect. Finally, the computation complexity is reduced to $\mathcal{O}(nh)$.

(3) Alternate Learning Strategy. Predicting sentiment based on real boundaries (i.e., Equation 4) can be more effective in learning the parameters in the sentiment prediction module, while leveraging boundary distributions (i.e., Equation 6) can enhance the capability of aspect detection. Therefore, different strategies have their own advantages, in this paper, we propose an alternate learning strategy to take the best of both tasks. Precisely, during training, we train a fixed ratio (τ) of samples based on Equation 4, while the others are learned using Equation 6. This strategy not only guides the sentiment prediction more properly but also eases the high computational cost caused by span enumeration.

3.4 Semantic Compatibility Mechanism

To predict the boundary probabilities of the aspects more accurately, we design a semantic compatibility mechanism. Intuitively, the words in an aspect should have some inherent connections, and they usually appear in the corpus simultaneously. To capture such information, we propose to explicitly learn the semantic compatibility between the words in an aspect. Formally, supposing the start and end boundaries of an aspect are s and e, respectively, then we need to measure the semantic compatibility of the words in this bound (s,e) to form a reasonable aspect. And the semantic compatibility score of the aspect is computed as,

$$m_{s,e} = \sigma\left(\boldsymbol{W}_6\text{ReLU}(\boldsymbol{W}_5 g(\boldsymbol{H}_{s\rightarrow e}) + \boldsymbol{b}_5) + \boldsymbol{b}_6\right), \tag{10}$$

where $\{\boldsymbol{W}_5, \boldsymbol{W}_6, \boldsymbol{b}_5, \boldsymbol{b}_6\}$ are trainable parameters. $\boldsymbol{H}_{s\rightarrow e}$ corresponds to the words in the aspect. g is

[2] the computational complexity is $\mathcal{O}(1)$ in $\mathcal{F}$.

Model	*Laptop*			*Restaurant*			*Tweets*		
	P	R	F1	P	R	F1	P	R	F1
SPJM (Zhou et al., 2019)	61.40	58.20	59.76	76.20	68.20	71.98	54.84	48.44	51.44
SPAN-pipeline (Hu et al., 2019)	69.46	66.72	68.06	76.14	73.34	74.92	60.72	55.02	57.69
SPAN-joint (Hu et al., 2019)	67.41	61.99	64.59	72.32	72.61	72.47	57.03	52.69	54.55
SPAN-collapsed (Hu et al., 2019)	50.08	47.32	48.66	63.63	53.04	57.85	51.89	45.05	48.11
SPRM (Lin and Yang, 2020)	68.66	68.77	68.72	77.78	80.60	79.17	60.25	58.79	59.45
S-AESC (Lv et al., 2021)	66.87	64.92	65.88	78.26	70.50	74.18	55.86	53.74	54.73
HI-ASA (Chen et al., 2022)	70.28	70.50	70.39	79.41	80.38	79.90	61.64	59.17	60.36
SIBP	75.18	66.40	**70.52**	84.45	76.15	**80.08**	66.76	55.51	**60.59**
%Improve.$^{+}$	6.97%	-5.82%	0.18%	6.35%	-5.27%	0.23%	8.31%	-6.19%	0.38%
p-value	3.39e-7*	2.55e-5*	0.042*	7.37e-5*	1.12e-5*	0.035*	1.97e-4*	3.49e-4*	0.012*

Table 1: The performance (Precision, Recall, and F1-Score) comparisons with different methods. Baseline results are retrieved from published papers. And the **bold** indicates the best performances among all the models.

$^{+}$ The improvements are calculated between HI-ASA and SIBP.

* It denotes that the corresponding improvement has passed the significant test at the significance level of 0.05.

a semantic aggregation function, there we specify it as the embedding combination from start to end position.

In the optimization process, we maximize the semantic compatibility scores between the real aspect boundaries, and simultaneously minimize that between the negative boundaries[3], that is,

$$\mathcal{J}_{sc} = \sum_{i=1}^{N}\sum_{j=1}^{l}\{\log m_{s_i^j, e_i^j} + \sum_{(x,y)\in O}\log(1-m_{x,y})\}, \quad (11)$$

where N is the number of sentences and l is the number of aspects in the sentence, O is the set of negative boundary pairs. In this equation, the parameters of $\{\boldsymbol{W}_5, \boldsymbol{W}_6, \boldsymbol{b}_5, \boldsymbol{b}_6\}$ are optimized to identify real aspects based on word semantics (i.e., $\boldsymbol{H}$), which provides complementary signals to the above index-based method.

During inference, For each candidate aspect (s, e), first, we judge whether it can compose an aspect based on γ (i.e., the sum score of start and end indexes). Then, we check whether this aspect pair is a semantically reasonable aspect via μ (see Equation 10). The specific details of our semantic-compatibility aspect extraction algorithm are shown in Appendix B.

3.5 Model Training

The overall loss is the weighted sum of the subtasks losses,

$$\mathcal{J} = \mathcal{J}_{ae} + \beta \cdot \mathcal{J}_{sp} + \gamma \cdot \mathcal{J}_{sc}, \quad (12)$$

where β and γ are task coefficients. We minimize the $\mathcal{J}$ and determine suitable parameters by grid search for information feedback mechanism during the experiment.

Aspect Extraction (F1)	*Laptop*	*Res*	*Tweets*
DE-CNN (Xu et al., 2018)	81.59	-	-
SPAN (Hu et al., 2019)	83.35	82.38	75.28
CL-BERT (Yang et al., 2020)	**85.61**	-	-
S-AESC (Lv et al., 2021)	85.19	84.20	76.04
- w/o DI	79.02	82.30	75.57
- w/o SC	81.66	84.45	75.34
- w/o SC & DI	81.11	85.01	74.25
SIBP	83.92	**85.87**	**76.54**
Sentiment Prediction (ACC)	*Laptop*	*Res*	*Tweets*
TNet (Li et al., 2018)	76.54	-	-
DMMN (Lin et al., 2019)	77.59	-	-
SPAN (Hu et al., 2019)	81.39	89.95	75.16
SPRM (Lin and Yang, 2020)	81.50	90.35	78.34
DGEDT (Tang et al., 2020)	76.80	-	74.80
DualGCN (Li et al., 2021)	81.80	-	77.40
T-GCN (Tian et al., 2021a)	81.97	-	78.03
- w/o DI	84.86	92.40	83.99
- w/o SC	84.70	92.44	84.11
- w/o SC & DI	83.28	92.01	82.99
SIBP	**85.96**	**92.49**	**84.13**
Joint Task (F1)	*Laptop*	*Res*	*Tweets*
- w/o DI	67.08	78.06	60.39
- w/o SC	68.93	79.05	60.22
- w/o SC & DI	68.30	77.54	58.82
SIBP	**70.52**	**80.08**	**60.59**

Table 2: The above is the performance (F1-score) comparisons with different methods on aspect extraction. The Middle is the performance (Accuracy) comparisons with different methods on sentiment prediction. And below is the performance (F1-score) comparisons with different methods on joint task.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct our experiments on three real-world datasets, which have been widely used in previous work (Hu et al., 2019; Lin and Yang, 2020; Chen et al., 2022). More details are shown in Appendix C.1.

Metrics. Please refer to Appendix C.2 for metric details.

Baselines. We compare our model with the fol-

[3] We sample the negative boundary pairs from all the boundary pairs that cannot form an aspect.

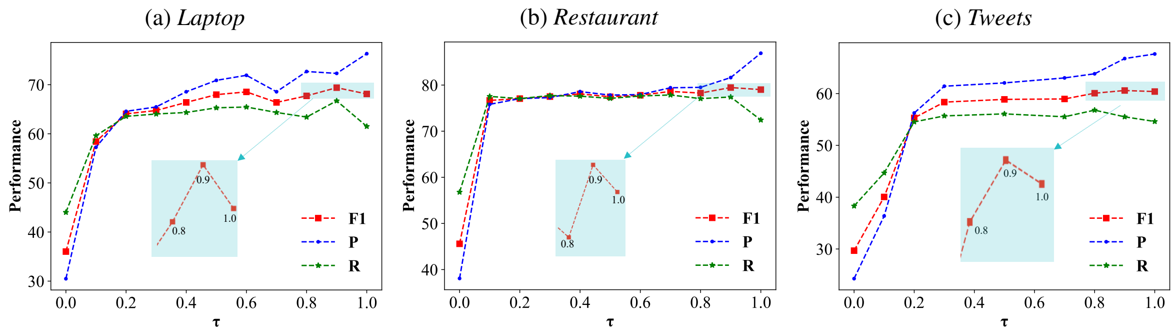


Figure 4: The performances (P, R, F1) of our SIBP w.r.t varying parameter τ on three datasets.

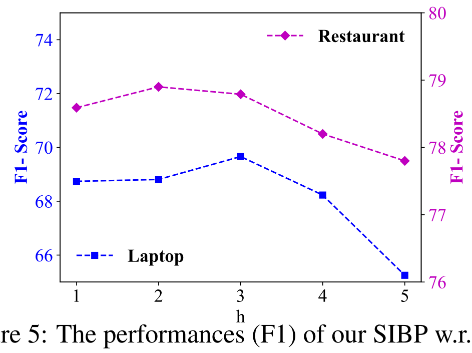


Figure 5: The performances (F1) of our SIBP w.r.t varying parameter h on *Restaurant* and *Laptop*, respectively.

lowing representative methods: **(1)** Joint aspect extraction and sentiment prediction, **(2)** Aspect extraction, **(3)** Sentiment Prediction.

(1) Joint aspect extraction and sentiment prediction method contain:
• **SPJM** (Zhou et al., 2019). • **SPAN-{pipeline, joint, collapsed}** (Hu et al., 2019). • **SPRM** (Lin and Yang, 2020) • **S-AESC** (Lv et al., 2021) • **HI-ASA** (Chen et al., 2022).

(2) Aspect extraction methods including • **DE-CNN** (Xu et al., 2018). • **CL-BERT** (Yang et al., 2020).

(3) Sentiment Prediction methods including • **TNet** (Li et al., 2018). • **DMMN** (Lin et al., 2019). • **TM** (Yadav et al., 2021). • **DGEDT** (Tang et al., 2020). • **DualGCN** (Li et al., 2021). • **T-GCN** (Tian et al., 2021a).

More baseline details are listed in Appendix C.3

Implementation Details. Implementation details are reported in Appendix C.4.

4.2 Main Results

The comparisons between our model and the baselines are presented in Table 1. Overall, our proposed model SIBP consistently achieves the best results against all the baselines. These observations indicate the carefully designed SIBP is capable of achieving better performances.

However, the recall performances of SIBP are lower than other baselines. This happens because we select the best model via the F1-score during optimization, which can maintain a good balance between precision and recall. Hence, we cannot guarantee that the precision and recall are at their peak when the model achieves the highest F1-score.

4.3 Ablation Study

In the above section, we have evaluated SIBP's overall performance. For studying the contributions of different model components, here, we go deeper into the model and conduct ablation studies. The results are shown in Table 2. More precisely, we compare SIBP with the following variants:

• - w/o DI: In this model, we only use the ground truth as the input of the sentiment prediction module (i.e., the ratio τ is set as 1).

• - w/o SC: In this model, we remove the semantic compatibility module (i.e., $\mathcal{J}_{sc}$).

The results in joint task show that when a certain module is removed, the model performance decreases, which indicates the indispensability of each module. This manifests that different model components are critical to the final performance.

In general, we can observe that SIBP outperforms baseline competitors on both tasks, which indicates the effectiveness of modeling the span-level associations between aspect extraction and sentiment classification.

4.4 Parameter Analysis

Effect of parameter τ. In our model, an important parameter is the ratio τ (i.e., **Remark.(3)**) when connecting different tasks. It measures how much we predict the sentiment based on the aspect boundary distributions. For the special cases, if $\tau = 0$, we make sentiment prediction completely based on the direct output from the aspect extraction module. If $\tau = 1$, the sentiment is predicted purely based on the ground truth boundaries. In this experiment, we tune the ratio τ in the range of 0 to 1, and the results are reported in Figure 4 based on the joint task.[4] We find on all datasets, the F1-score increases rapidly at first, then tends to be stable, and

[4] It should be noted that $\tau = 1$ is equal to the model of -w/o DI in the above ablation studies.

Sentences	Predictions of HI-ASA	SIBP
1. All the money went into the **[interior decoration]**$_+$, none of it went to the **[chefs]**$_-$.	[interior decoration]$_+$(✓) [chefs]$_+$(✗)	[interior decoration]$_+$(✓) [chefs]$_-$(✓)
2. After **[dinner]**$_-$, take your date to the HUGE **[dance floor]**$_+$, probably one of the biggest you'll see in NY.	[dinner]$_+$(✗) [dance floor]$_+$(✓)	[dinner]$_-$(✓) [dance floor]$_+$(✓)
3. **[Lunch]**$_+$ came with **[pickels and slaw]**$_-$, no extra charge.	[lunch]$_+$(✓) [pickels]$_-$(✗) [slam]$_-$(✗)	[lunch]$_+$(✓) [pickels and slaw]$_-$(✓)
4. You will obtain a gift if you buy the separate **[RAM memory]**$_+$.	[separate RAM memory]$_+$(✗)	[RAM memory]$_+$(✓)

Table 3: Outputs of different models. The extracted aspects are wrapped in brackets with the predicted polarities given as subscripts. Correct and incorrect predictions are marked with ✓and ✗, "+/-/0" denote the positive, negative, and neutral sentiment polarities.

finally falls down from $\tau = 0.9$ to $\tau = 1$. The reason lies in that when τ is small, the model cannot obtain enough supervised information from ground truth, resulting in poor performance. And when τ is very large, the model cannot take into account the valuable interactions between the two tasks, leading to unsatisfactory performance. We observe that the best performances are usually achieved when τ is moderate (i.e., 0.9) on three datasets, which implies that neither the ground truth boundary nor distributional boundary is dominantly superior, and a mixture of them can be more favorable.

Effect of parameter h. In this section, we investigate the influence of h, which controls the aspect's length in Equation 9. Theoretically, when h is large, the model parameters are huge and difficult to converge. Conversely, when h is very small, it is not easy to extract correct aspects. Considering the length of an aspect term is usually not large, we search h in the range of $\{1, 2, 3, 4, 5\}$ in our experiment. We provide the F1-scores on *Restaurant* and *Laptop* in Figure 5. As h increases, the performances gradually rise to a maximum and then start to fall, indicating that a moderate h (2 or 3) can not only reduce the amount of parameters but also ensure extracting correct aspects.

4.5 Case study

To provide more insights into our model's superior performance, we present some case studies in a qualitative manner. Table 3 shows that:

• In the first two cases, since the associations between aspect extraction and sentiment prediction are weak, the aspect detection and sentiment in the HI-ASA model are not completely correct. For instance, it extracts the correct aspect "dinner", but predicts inaccurate sentiment with "dinner" in the 2nd case. However, in SIBP, the sentiment information can be backpropagated into the boundary detection process, which can enhance the correlations and therefore predict the aspect and sentiment correctly.

• By modeling the semantic compatibility, SIBP can extract more accurate aspects. For example, in the 4th case, the word "separate" has a higher probability to become the start of an aspect as compared to the word "RAM", while from the semantic perspective, "separate" should not be mixed into the aspect of "RAM memory", which has been successfully discovered by our model. Similar phenomena can also be observed in the 3rd case.

5 Conclusion and Future Work

In this paper, we studied the problem of joint aspect extraction and sentiment prediction. Different from previous approaches that model the inter-task correlations replying on sentence-level features, we proposed to model the span-level interactions with boundary probabilities for the task of targeted sentiment analysis. In order to take the best of both tasks, we devised an alternate learning strategy to improve the performance during training. Moreover, we combined the position and semantic information to detect the boundaries of an aspect. Extensive experiments on three real-world datasets demonstrated the model's superiority.

This paper actually makes the first step towards modeling span-level interactions for targeted sentiment analysis. However, there is still much room to be improved. For example, the position information and semantic information are scattered in a sentence, which will affect the performance of joint aspect-sentiment extraction. In future work, we may introduce graph structure to model the position and semantic information in a more integrated manner. Besides, we may also leverage some data augmentation technologies to solve the sparse problem in targeted sentiment analysis.

References

Liu Bing. 2012. Sentiment analysis and opinion mining. In *Synthesis Lectures on Human Language Technologies 5.1 (2012): 1-167.*

Guimin Chen, Yuanhe Tian, and Yan Song. 2020. Joint aspect extraction and sentiment analysis with directional graph convolutional networks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 272–279.

Wei Chen, Jinglong Du, Zhao Zhang, Fuzhen Zhuang, and Zhongshi He. 2022. A hierarchical interactive network for joint span-based aspect-sentiment analysis. *arXiv preprint arXiv:2208.11283*.

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.

Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheng Li, and Yiwei Lv. 2019. Open-domain targeted sentiment analysis via span-based extraction and classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 537–546.

Niklas Jakob and Iryna Gurevych. 2010. Extracting opinion targets in a single and cross-domain setting with conditional random fields. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1035–1045.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 151–160.

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

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Hao Li and Wei Lu. 2017. Learning latent sentiment scopes for entity-level sentiment analysis. In *AAAI*, pages 3482–3489.

Ruifan Li, Hao Chen, Fangxiang Feng, Zhanyu Ma, Xiaojie Wang, and Eduard Hovy. 2021. Dual graph convolutional networks for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6319–6329.

Xin Li, Lidong Bing, Wai Lam, and Bei Shi. 2018. Transformation networks for target-oriented sentiment classification. *arXiv preprint arXiv:1805.01086*.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019. A unified model for opinion target extraction and target sentiment prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6714–6721.

Peiqin Lin and Meng Yang. 2020. A shared-private representation model with coarse-to-fine extraction for target sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4280–4289.

Peiqin Lin, Meng Yang, and Jianhuang Lai. 2019. Deep mask memory network with semantic dependency and context moment for aspect level sentiment classification. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 5088–5094. AAAI Press.

Huaishao Luo, Tianrui Li, Bing Liu, and Junbo Zhang. 2019. Doer: Dual cross-shared rnn for aspect term-polarity co-extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 591–601.

Yanxia Lv, Fangna Wei, Ying Zheng, Cong Wang, Cong Wan, and Cuirong Wang. 2021. A span-based model for aspect terms extraction and aspect sentiment classification. *Neural Computing and Applications*, 33(8):3769–3779.

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

Margaret Mitchell, Jacqui Aguilar, Theresa Wilson, and Benjamin Van Durme. 2013. Open domain targeted sentiment. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1654.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Maria Pontiki, Dimitrios 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.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, AL-Smadi Mohammad, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 19–30.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495.

Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2016. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 108:42–49.

Hao Tang, Donghong Ji, Chenliang Li, and Qiji Zhou. 2020. Dependency graph enhanced dual-transformer structure for aspect-based sentiment classification. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6578–6588.

Yuanhe Tian, Guimin Chen, and Yan Song. 2021a. Aspect-based sentiment analysis with type-aware graph convolutional networks and layer ensemble. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2910–2922.

Yuanhe Tian, Guimin Chen, and Yan Song. 2021b. Enhancing aspect-level sentiment analysis with word dependencies. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3726–3739.

Xinyi Wang, Guangluan Xu, Zequn Zhang, Li Jin, and Xian Sun. 2021. End-to-end aspect-based sentiment analysis with hierarchical multi-task learning. *Neurocomputing*, 455:178–188.

Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. 2018. Double embeddings and cnn-based sequence labeling for aspect extraction. *arXiv preprint arXiv:1805.04601*.

Rohan Kumar Yadav, Lei Jiao, Ole-Christoffer Granmo, and Morten Goodwin. 2021. Human-level interpretable learning for aspect-based sentiment analysis. In *The thirty-fifth AAAI conference on artificial intelligence (AAAI-21). AAAI*.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021. A unified generative framework for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2416–2429.

Yunyi Yang, Kun Li, Xiaojun Quan, Weizhou Shen, and Qinliang Su. 2020. Constituency lattice encoding for aspect term extraction. In *Proceedings of the 28th international conference on computational linguistics*, pages 844–855.

Meishan Zhang, Yue Zhang, and Duy Tin Vo. 2015. Neural networks for open domain targeted sentiment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 612–621.

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*.

Yan Zhou, Longtao Huang, Tao Guo, Jizhong Han, and Songlin Hu. 2019. A span-based joint model for opinion target extraction and target sentiment classification. In *IJCAI*, pages 5485–5491.

Appendix

A Proof

In this section, we prove that the optimization objective of targeted sentiment analysis does not change after partitioning the training dataset.

First, when there are multiple aspects in a sentence, the optimization objective can be denoted as,

$$\begin{aligned}
\mathcal{J}_1 &= \mathcal{J}_{ae} + \mathcal{J}_{sc} + \mathcal{J}_{sp} \\
&= -\sum_{i=1}^{N}\sum_{i=1}^{m}\left\{\boldsymbol{p}_{s,i,j}^{T}\log\left(\hat{\boldsymbol{p}}_{s,i,j}\right) + \boldsymbol{p}_{e,i,j}^{T}\log\left(\hat{\boldsymbol{p}}_{e,i,j}\right)\right\} \\
&\quad + \sum_{i=1}^{N}\sum_{j=1}^{l}\{\log m_{s_i^j} e_i^j + \sum_{(x,y)\in O}\log\left(1 - m_{x,y}\right)\} \\
&\quad - \sum_{i=1}^{N}\sum_{j=1}^{l}\sum_{k=1}^{t} y_{i,j,k}\log\hat{y}_{i,j,k} \\
&= -\sum_{i=1}^{N}\left(\sum_{i=1}^{m}\left\{\boldsymbol{p}_{s,i,j}^{T}\log\left(\hat{\boldsymbol{p}}_{s,i,j}\right) + \boldsymbol{p}_{e,i,j}^{T}\log\left(\hat{\boldsymbol{p}}_{e,i,j}\right)\right\}\right. \\
&\quad - \sum_{j=1}^{l}\{\log m_{s_i^j,e_i^j} + \sum_{(x,y)\in O}\log\left(1 - m_{x,y}\right)\} \\
&\quad \left. + \sum_{j=1}^{l}\sum_{k=1}^{t} y_{i,j,k}\log\hat{y}_{i,j,k}\right) \\
&= -\sum_{i=1}^{N}\left(\sum_{i=1}^{m}\left\{\boldsymbol{p}_{s,i,j}^{T}\log\left(\hat{\boldsymbol{p}}_{s,i,j}\right) + \boldsymbol{p}_{e,i,j}^{T}\log\left(\hat{\boldsymbol{p}}_{e,i,j}\right)\right\}\right. \\
&\quad - \sum_{j=1}^{m}\{\log m_{s_i^j,e_i^j} + \sum_{(x,y)\in O}\log\left(1 - \hat{m}_{x,y}\right)\} \\
&\quad \left. + \sum_{j=1}^{m}\sum_{k=1}^{t} y_{i,j,k}\log\hat{y}_{i,j,k}\right) \\
&= -\sum_{i=1}^{N}\sum_{i=1}^{m}\left(\left\{\boldsymbol{p}_{s,i,j}^{T}\log\left(\hat{\boldsymbol{p}}_{s,i,j}\right) + \boldsymbol{p}_{e,i,j}^{T}\log\left(\hat{\boldsymbol{p}}_{e,i,j}\right)\right\}\right. \\
&\quad - \{\log m_{s_i^j,e_i^j} + \sum_{(x,y)\in O}\log\left(1 - \hat{m}_{x,y}\right)\} \\
&\quad \left. + \sum_{k=1}^{t} y_{i,j,k}\log\hat{y}_{i,j,k}\right)
\end{aligned} \tag{13}$$

where N is the number of sentences in training dataset, m is the sentence's length, and l is the aspect's number in each sentence. With the above derivation, we can observe that the targeted sentiment analysis can be viewed as an optimization of a single aspect. In other words, this is independent of the number of aspects in a sentence.

To further test our idea, we derive the case that there is only one aspect per sentence, the optimization goal is,

$$\begin{aligned}
\mathcal{J}_2 &= \mathcal{J}_{ae} + \mathcal{J}_{sc} + \mathcal{J}_{sp} \\
&= -\sum_{i=1}^{M}\sum_{i=1}^{m}\left(\left\{\boldsymbol{p}_{s,i,j}^{T}\log\left(\hat{\boldsymbol{p}}_{s,i,j}\right) + \boldsymbol{p}_{e,i,j}^{T}\log\left(\hat{\boldsymbol{p}}_{e,i,j}\right)\right\}\right. \\
&\quad - \{\log m_{s_i^j,e_i^j} + \sum_{(x,y)\in O}\log\left(1 - \hat{m}_{x,y}\right)\} \\
&\quad \left. + \sum_{k=1}^{t} y_{i,j,k}\log\hat{y}_{i,j,k}\right),
\end{aligned} \tag{14}$$

where M is the number of sentence in new training dataset, i,e., $M = \sum_{i=1}^{N} l_i$.

We can observe that $\mathcal{J}_1$ and $\mathcal{J}_2$ are linearly dependent, thereby, the two optimization objectives are consistent.

B Semantic-compatibility Aspect Extraction Algorithm

See algorithm 1.

Algorithm 1 Overall Process

```
Input: $g_s$ , $g_e$ , $k$ , $\gamma$ , $\mu$ .
    $g_s$ : the score of start position
    $g_e$ : the score of end position
    $k$ : the maximum number of extracted aspects
    $\gamma$ : sum score threshold
    $\mu$ : compatibility score threshold
Output: $\mathcal{M}$ denotes the aspects set
Initialize the extracted aspect pair $\mathcal{M}$=$\varnothing$
Get the top-$k$ start indexes $S$ from $g_s$
Get the top-$k$ end indexes $E$ from $g_e$
for $i$ in $S$ do
    for $j$ in $E$ do
        if $i \leq j$ then
            if $g_{s_i} + g_{e_j} \geq \gamma$ and $m_{s,e} \geq \mu$ then
                $\mathbf{u} = j - i$
                $\mathbf{r} = g_{s_i} + g_{e_j} - \mathbf{u}$
                $\mathbf{t} = \{(i, j)\}$
                $\mathcal{M} \leftarrow \mathcal{M} \cup \mathbf{t}$
Sort $\mathcal{M}$ in ascending order by the value of $\mathbf{r}$
for $k$ in size($\mathcal{M}$) do
    for $q$ in size($\mathcal{M}$) do
        if $\mathcal{M}_k \cap \mathcal{M}_q \neq \varnothing$ then
            Delete $\mathcal{M}_k$
Return $\mathcal{M}$
```

C Experimental Setup

C.1 Datasets

We conduct our experiments on three real-world datasets, which have been widely used in previous

Dataset	#Sentences	#Aspects	#+	#-	#0
Laptop	1869	2936	1326	900	620
Restaurant	3900	6603	4134	1538	931
Tweets	2350	3243	703	274	2266

Table 4: Statistics of three datasets. "+/-/0" denote the positive, negative, and neutral sentiment polarities.

work and the statistics are shown in Table 4. Precisely, *Laptop* contains costumer reviews in electronic product domain, which is collected from SemEval Challenge 2014 (Pontiki et al., 2014). *Restaurant* is the reviews set of the restaurant domain from SemEval2014, SemEval2015 and SemEval2016 (Pontiki et al., 2014, 2015, 2016). *Tweets* is built by (Mitchell et al., 2013), which is composed of twitter posts from different users. These datasets come from different domains with various characters, for example, *Restaurant* contains more positive or negative sentiments, while there are relatively fewer neutral comments. In contrast, the number of neutral sentiments in *Tweets* is much larger than that of the positive or negative polarities. With these datasets, our model can be fairly evaluated under different settings.

C.2 Metrics

In the experiments, the commonly used metrics including Precision (P), Recall (R), and F1-Score (F1) are selected to evaluate our model. For aspect extraction, we also use F1-Score to evaluate our model. And we adopt accuracy (ACC) as the metric in sentiment prediction. In specific, a predicted target is correct only if it exactly matches the gold targets and the corresponding polarity. Since we use exactly the same datasets and metrics as (Hu et al., 2019), we directly copy the baseline results from it.

C.3 Baselines

We compare our model with the following representative methods: **(1)** Joint aspect extracting and sentiment predicting, **(2)** Aspect extracting, **(3)** Sentiment Predicting.

(1) Joint aspect extraction and sentiment prediction methods contain:

- **SPJM** (Zhou et al., 2019): The model firstly searches directly to determine the span of the target and then performs sentiment polarity classification.
- **SPAN-{pipeline, joint, collapsed}** (Hu et al., 2019): A span-based model employing a BERT encoder and a multi-target span decoder to extract aspects and corresponding sentiments.
- **SPRM** (Lin and Yang, 2020): It adopts the private and share representation for the joint task to capture correlations between the two tasks, which is an improvement over SPAN (Hu et al., 2019).
- **S-AESC** (Lv et al., 2021): A dual gated recurrent units and an interaction layer are used jointly to generate the aspects and sentiments.
- **HI-ASA** (Chen et al., 2022): The current state-of-the-art model for span-based targeted sentiment analysis, which proposes a hierarchical interactive network that takes explicit cooperation from aspect extraction and sentiment predicting to exact the aspects and sentiments.

(2) Aspect extraction methods including:

- **DE-CNN** (Xu et al., 2018): It utilizes a double embedding with CNNs to extract aspect words.
- **CL-BERT** (Yang et al., 2020): The current state-of-the-art model for aspect extraction, this work incorporates the syntactic information in neural network models.

(3) Sentiment Prediction methods including:

- **TNet** (Li et al., 2018): A classification model that uses a multi-layer context-preserving network as the feature extractor.
- **DMMN** (Lin et al., 2019): It integrates the information of semantic dependency and inter-aspect relation into memory network.
- **DGEDT** (Tang et al., 2020): This work proposes a dependency graph enhanced dual-transformer network, which joint considers the flat representation learned from Transformer and graph-based representations learned from the corresponding dependency graph.
- **DualGCN** (Li et al., 2021): It designs a dual graph convolution network that considers the complementarity of syntax structures and semantic correlations simultaneously.
- **T-GCN** (Tian et al., 2021a): A state-of-the-art method that explicitly uses dependency types for ABSA with type-aware graph convolution network.

C.4 Implementation Details

Following previous works (Hu et al., 2019; Lin and Yang, 2020; Chen et al., 2022), we split the training and test sets for each dataset. As there is no train-test split in *Tweets*, we report the ten-fold cross-validation results for it. In the model, we adopt the BERT-large model as the backbone network, which consists of 24 Transformer blocks with 16 self-attention heads as the encoding layer. The maximum length of aspect (i.e., h in Equation 9) is in the range of $\{1, 2, 3, 4, 5\}$. The ratio of distributional input samples (i.e., τ in **Remark.(3)**) is determined in the range of 0 to 1. our model is optimized based on Adam optimizer (Kingma and Ba, 2014), and the learning rate is searched in $\{2e\text{-}5, 2e\text{-}4, 2e\text{-}3\}$. Moreover, all methods were implemented using the PyTorch framework and trained on Nvidia GeForce Titan RTX 3090 GPUs.