HOSSemEval-EB23: A Robust Dataset for Aspect-Based Sentiment Analysis of Hospitality Reviews

Tram T. Doan[1,2], Thuan Q. Tran[1,2], Dat T. Le[1,2], Anh H. Tran[1,2], An T. Nguyen[1,2,3], An-Tran Hoai-Le[1,2,3], Tran-Tung Doan-Nguyen[1,2,3], Son T. Huynh[1,2,3], Binh T. Nguyen[1,2,3*]

[1]University of Science, Ho Chi Minh City, 700000, Vietnam.
[2]Vietnam National University, Ho Chi Minh City, 700000, Vietnam.
[3]AISIA Research Lab.

*Corresponding author(s). E-mail(s): ngtbinh@hcmus.edu.vn;

Abstract

Aspect-Based Sentiment Analysis (ABSA), also known as fine-grained opinion mining, is a crucial task focused on understanding the sentiment expressed in a text with respect to specific aspects. ABSA has become increasingly vital in the digital era due to its significant contributions in extracting insights and facilitating decision-making processes. As a result, ABSA has garnered considerable attention and has been widely applied across diverse domains. The evolution of ABSA is showcased through its expansion into various domains and the adoption of advanced machine-learning techniques. This paper contributes to the ABSA field by introducing a novel dataset specifically tailored to the hospitality domain. The dataset serves as a valuable resource for benchmarking the performance of sentiment analysis models. By providing context for sentiment expression, the dataset offers detailed insights into various aspects such as facilities, amenities, services, overall experience, brand image, and customer loyalty. This rich information enables more nuanced analysis and enhances our understanding of customer sentiments within the hospitality industry.

Keywords: ABSA, Sentiment Analysis, Hospitality Reviews

1 Introduction

The digital landscape has revolutionized how consumers share their experiences with products, services, and travel destinations. Online review platforms like TripAdvisor and Booking.com offer a treasure trove of information for both travelers seeking informed decisions and businesses striving to improve customer satisfaction and service quality. This information is particularly valuable due to the rise of Sentiment Analysis (SA) and Aspect-Based Sentiment Analysis (ABSA). SA and ABSA are crucial tools for extracting opinions from the text about specific entities and their corresponding aspects (Liu, 2012) [1]. Think of it as dissecting customer feedback at a granular level, uncovering overall sentiment and specific opinions towards various aspects like hotel amenities or staff service. This provides businesses with valuable insights into customer experiences, enabling them to pinpoint areas for improvement and ultimately refine their offerings.

ABSA tasks typically involve two key subtasks: aspect extraction and aspect-based sentiment classification. By combining these, we can identify an aspect-sentiment pair, where the first element represents the specific aspect mentioned (e.g., cleanliness), and the second reflects the corresponding sentiment (e.g., positive, negative). (Hu and Liu, 2004; Qiu et al., 2011) [2, 3] demonstrate the vital link between aspect terms and opinion terms in improving ABSA accuracy. Early research in ABSA focused on single-factor prediction, such as extracting specific aspect terms (Liu et al., 2015; Xu et al., 2018)[4, 5] or discovering broader aspect categories (Zhou et al., 2015) [6]. Others concentrated on classifying sentiment based on identified aspects (Ruder et al., 2016; Hu et al., 2019a) [7, 8] or even considering only single aspect terms (Huang and Carley, 2018) [9]. Modern advancements, however, explore extracting multiple emotion factors simultaneously (Zhang et al., 2021) [10]. These studies often rely on existing datasets (Pontiki et al., 2014, 2015, 2016) [11–13] or ones augmented with additional sentiment annotations (Fan et al., 2019; Li et al., 2019a; Xu et al., 2020; Wan et al., 2020) [14–16]. While valuable, these datasets often have limitations. Many are monolingual, focusing on analyzing sentiment at the sentence or paragraph level, neglecting the richness of individual sub-sentences within longer reviews. This makes distinguishing nuanced opinions about multiple aspects challenging, especially in complex paragraphs.

This paper introduces HOSSemEval-EB23, a novel dataset constructed from customer reviews within the hospitality domain. It includes annotations for target aspect category polarity, facilitating ABSA tasks. This dataset not only provides context for sentiment expression but also uncovers rich information about specific aspects like facilities, amenities, services, overall experience, brand image, and even customer loyalty. This enhances practicality, allowing businesses to gain a clearer understanding of the individual factors influencing customer evaluations and perceptions. Our dataset further includes information on the specificity of opinions by associating aspect terms with both aspect categories and sentiment polarity. This diversity and accuracy minimize challenges for models dealing with vague or ambiguous opinions.

The contribution of our work can be described as follows:

(a) We introduce a new dataset of customer reviews in the hospitality domain, providing a valuable resource for future research on the ABSA task.

(b) Leveraging this novel dataset to substantially address gaps in existing ABSA datasets and enrich the diversity of topics explored in ABSA tasks, especially within the hotel domain.
(c) This dataset can serve as a benchmark for evaluating the performance of existing methods or developing new techniques in sentiment analysis or related fields.

Our paper is structured as follows: Firstly, section [2] provides an overview of datasets and various methodological approaches used in ABSA tasks. Then, detailed information regarding the new dataset annotation and characteristics can be found in Section [3]. Additionally, Section [4] outlines the main evaluation methods, while Section [5] presents the details of experiment results analysis related to evaluating our proposed dataset. Finally, the paper concludes with a conclusion on future directions.

2 Related Work

ABSA has garnered significant attention in recent years due to its importance in understanding the nuanced sentiment expressed in textual data. Studies in ABSA have flourished, encompassing the development and expansion of methodologies, datasets, models, evaluation metrics, applications, and emerging trends. The advancement of ABSA has been explored across various domains, including consumer electronics, movies, and restaurants, leveraging diverse datasets and methodologies. In 2014, Pontiki et al. introduced datasets featuring manually annotated reviews in the domains of restaurants and laptops, named SemEval-2014 Task 4 (SE-ABSA14) [11]. This initiative garnered significant interest, receiving 163 submissions from 32 teams that experimented with a variety of features. A year later, SemEval-2015 Task 12 (SE-ABSA15) [12] expanded upon SE-ABSA14, continuing to provide manually annotated reviews across three domains (restaurants, laptops, and hotels), along with a standardized evaluation procedure. This iteration attracted 93 submissions from 16 teams. In the subsequent year, SemEval 2016 (SE-ABSA16 task) [13] maintained the focus on ABSA, serving as a continuation of the respective tasks from 2014 and 2015. The task attracted 245 submissions from 29 teams, indicating continued interest and participation in ABSA research.

SemEval-2014, SemEval-2015, and SemEval-2016 have demonstrated their significance in the field of ABSA by attracting substantial research attention, serving as a foundation for the development and expansion of studies, and making notable contributions to the field. They have become benchmark datasets to evaluate the effectiveness of methods in this domain. Many studies have utilized these established datasets to predict sentiment polarities, surpassing outstanding achievements. In 2020, Wan [17] proposed the TAS method for identifying aspect terms, aspect categories, and sentiment polarity in textual data. By leveraging datasets from SemEval-2015 and SemEval-2016, Wan's proposed method achieved high performance in detecting aspect term - aspect category - sentiment polarity triples, including implicit aspect term cases. This method also plays a crucial role in evaluating the significance of datasets within the ABSA domain. However, most studies [15, 17–19] have primarily focused on jointly predicting multiple elements in textual data, ignoring the rich label semantics inherent in ABSA problems. Recognizing the potential benefits of incorporating label

semantics for the joint extraction of multiple sentiment elements, Zhang [10] proposed a new approach. This approach tackled various ABSA tasks within a unified generative framework, employing two paradigms—annotation-style and extraction-style modeling. These paradigms were designed to facilitate the training process by framing each ABSA task as a text-generation problem. Zhang and his team conducted experiments on four ABSA tasks across multiple benchmark datasets to evaluate the effectiveness of their approach. There was no doubt that Wan and Zhang's approach was popular in ABSA. However, these tasks only attempt partial prediction instead of identifying all four sentiment elements in one shot. To address this limitation, Zhang [20] and his team introduced the ASQP method to predict all quadruples (aspect category, aspect term, opinion term, sentiment polarity) for a given opinionated sentence.

3 Dataset

3.1 Data Collection and Annotation

We conducted data collection from Booking.com[1], covering the period from 2020 to 2023. The collected data were in various languages, including English (EN), Vietnamese (VN), Chinese (CH), and Korean (KR), among others. However, our analysis was exclusively focused on reviews written in English. The dataset, relevant to the hotel domain, includes customer ratings, the locations they stayed at, their general information, and their reviews.

After the data collection phase, we initiated data preprocessing. Post-preprocessing, we embarked on the annotation process using the Label Studio platform[2]. A total of 7,012 data points were annotated, comprising 60,282 sentences, 30,141 aspects, and 30,141 sentiment counts. The annotation was guided by specific keyword definitions vital for analyzing the hotel domain:

Aspect Keywords:

- *Facility:* Represents aspects such as public/room equipment, cleanliness, hotel decor, interior design, balconies, and pool areas.
- *Amenity:* Encompasses public services like parking, spas, restaurants, souvenirs, and nearby destinations, including payment options, hotel location, general safety, and security.
- *Service:* Pertains to restaurant service, overall food quality, and staff behavior, including their helpfulness, friendliness, and product knowledge.
- *Experience:* Relates to the hotel's unique characteristics, like the local surroundings, atmosphere, overall view, and the level of relaxation it offers.
- *Branding:* Reflects the overall customer satisfaction compared to the hotel's rating, including service quality relative to expectations or information provided.
- *Loyalty:* Indicates the likelihood of guests revisiting and recommending the hotel to others.

Sentiment Keywords:

[1] https://www.booking.com
[2] https://labelstud.io/

- *Positive:* Assigned when a customer is completely satisfied with the mentioned aspect.
- *Negative:* Assigned when a customer expresses dissatisfaction or discomfort regarding the aspect.
- *Neutral:* Assigned when a customer is generally satisfied but notes minor areas for improvement.

Table 1 below describes three examples of how data were annotated, which defined aspect and sentiment.

Table 1: Three examples of Data Annotation

ID: booking_hcm_24572 **Sentence:** Clean room with huge comfortable bed. The staff was very helpful. I recommend this place.		
Aspect Term	**Aspect Category**	**Sentiment Polarity**
Clean room with huge comfortable bed	Facility	Positive
The staff was very helpful and kind	Service	Positive
I recommend this place	Loyalty	Positive

ID: booking_hn_60855 **Sentence:** The toilet is very small, lock of the door is break. It's noisy around. The host is not friendly.		
Aspect Term	**Aspect Category**	**Sentiment Polarity**
The toilet is very small	Facility	Negative
lock of the door is break	Facility	Negative
It's noisy around	Experience	Negative
The host is not friendly	Service	Negative

ID: booking_hcm_88697 **Sentence:** Staff were all amazing! The pool was a nice respite from the busyness of the city. Wish the fitness room was a little bigger and better stocked.		
Aspect Term	**Aspect Category**	**Sentiment Polarity**
Staff were all amazing	Service	Positive
The pool was a nice respite from the busyness of the city	Facility	Positive
Wish the fitness room was a little bigger and better stocked	Facility	Neutral

3.2 Preprocessing

Our data originates from customer reviews of Vietnamese restaurants and hotels harvested through Booking.com, a prominent online travel platform. To prepare this Booking.com dataset for analysis, we implemented a series of crucial cleansing steps to refine the data's suitability. Firstly, duplicate entries were eliminated to guarantee data uniqueness. Secondly, reviews lacking ratings were excluded. Thirdly, reviews without textual content were discarded to focus exclusively on text-based feedback. Furthermore, all emojis were removed for streamlined textual analysis. Additionally, a language detection process ensured linguistic consistency by retaining only reviews in English. Finally, each review was assigned a unique reference ID to facilitate future data handling and retrieval. The ensuing sections will provide a comprehensive breakdown of the following key steps:

- Duplicate Removal: To guarantee data uniqueness, all duplicate entries are identified and eliminated. Typically, review identifiers (e.g., Review ID) serve as the basis for identifying duplicates
- Filtering Out Non-Rated Reviews: Reviews lacking explicit ratings or rating-related attributes are identified and filtered out. Additionally, non-review entries, such as metadata or irrelevant text, are excluded to enhance dataset cohesiveness
- Emoji Removal: Use regular expressions or dedicated libraries to detect and remove emojis from text. Alternatively, replace emojis with a placeholder or remove them entirely, which is not necessary.
- Language Detection and Filtering: Reviews written in languages other than the target language (e.g., English for our current goal) are filtered out using language detection techniques. This ensures linguistic consistency within the dataset.
- Assignment of Reference IDs: Each review is assigned a unique reference identifier (e.g., sequential numbers, alphanumeric strings). These identifiers facilitate seamless data handling and retrieval throughout the analysis pipeline, ensuring data traceability and organization.

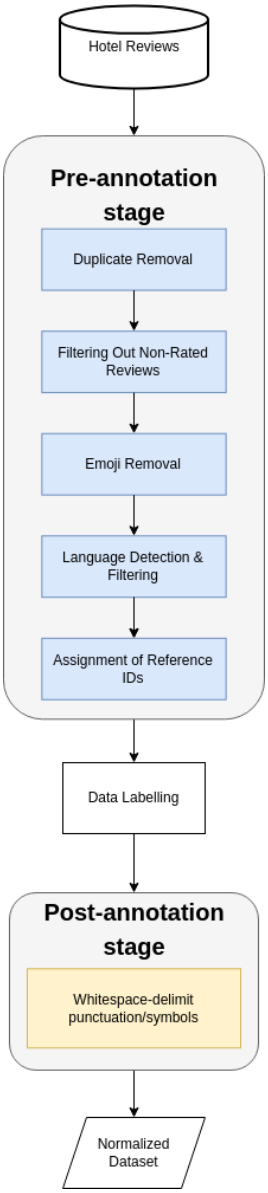

Fig. 1: Flowchart of data preprocessing

To make human labeling easier, we processed the raw data as follows to reduce confusion (Figure 1). The labeling was based on two criteria mentioned in the data

collection section. After completing the labeling, we isolated punctuation or special symbols by adding whitespace before and after them. We treated these as words. As with other NLP processing steps, we lowercase the entire text and did not remove stopwords. Stopwords are valuable because they can contain important information about the meaning of a sentence. If we removed stopwords, we could lose this information.
To guarantee the data's suitability for evaluating different model types, we propose two distinct normalization approaches tailored to specific problem requirements. One method is designed for generative models, while the other caters to TAS-Transformer models.

Generative model:

- The input data undergoes a dedicated preprocessing pipeline to support efficient training and inference within the generative model framework for ABSA, particularly methodologies like GAS and ASQP. This pipeline aligns the data with the specific format employed by the benchmark rest15, 16 datasets. This format is represented as:
  Sentence. #### [[aspect term*$_1$*, aspect category*$_1$*, sentiment polarity*$_1$*], [aspect term*$_2$*, aspect category*$_2$*, sentiment polarity*$_2$*], ..., [aspect term*$_n$*, aspect category*$_n$*, sentiment polarity*$_n$ *]]
  For example: Lovely helpful staff and very clean and quiet. #### [['lovely helpful staff', 'service', 'positive'], ['very clean and quiet', 'facility', 'positive']
- We split the sentence into a sequence of tokens, and we identified the aspect terms, aspect categories, and sentiment polarities. Occasionally, due to annotation inconsistencies or algorithm errors, the extracted aspect category and sentiment polarity might be incorrectly ordered. To address this, a final step verifies the last element in the sequence. If it does not belong to the set aspect term, aspect category, sentiment polarity, then the order of the last two elements is swapped.

TAS-Transformers model:

- To train the proposed model, the training set must undergo preprocessing to generate a tuple (S, a, p, f, T) for each unique combination of S, a, and p. Here, S represents a sentence from the training set, a is an aspect category found within the training set, p is a sentiment polarity identified in the training set, f denotes a binary "yes/no" label, and T is a sequence of labels formatted according to either the BIO or TO tagging schemes. Here is an example utilizing both the BIO and TO tagging schemes.
  Using BIO Tag: *(the room is very clean ., facility, positive, yes, B I I I I O)*
  Using TO Tag: *(the room is very clean ., facility, positive, yes, T T T T T O)*

4 Methodology

In this section, we describe distinct approaches used to evaluate our proposed dataset.

4.1 TAS-Transformer Models

The study by Wan et al. (2020) [17] introduces an innovative neural-based approach to detect all triplets (aspect term, aspect category, sentiment polarity) in a given sentence. This approach addresses the challenge of analyzing aspect-level sentiment, where the sentiment often depends on both the aspect term and the aspect category. The method divides the task into two interrelated subproblems based on pairs of (aspect categories and sentiment polarity). The first subproblem focuses on identifying the presence of aspect terms within these aspect-sentiment pairs, functioning as a binary text classification challenge. This involves determining whether any aspect term exists for a specific combination of aspect category and sentiment polarity. If such a combination is present, the second subproblem, a sequence labeling task, aims to pinpoint the exact aspect term. For example, in the sentence "*Very modern and spacious room*" with a (Facility, positive) pair, the first subtask would confirm the existence of an aspect term in this aspect category - sentiment polarity pair, and the sequence labeling model would then identify "*Very modern and spacious room*" as the corresponding aspect term, resulting in the output of a triplet (Very modern and spacious room, facility, positive). Conversely, with a (service, positive) pair in the same sentence, the model would predict that no aspect term exists in this aspect category-sentiment polarity pair. The second subproblem can be simplified to a sequence labeling task, employing either the BIO tagging scheme or the TO tagging scheme. In the BIO scheme, "B" (resp. "I") signifies the beginning (resp. internal) of an aspect term, while "T" denotes a word within an aspect term, and "O" represents a word outside any aspect term.

Both subproblems are solved by a single neural model that leverages the capabilities of the pre-trained BERT language model (Devlin et al.,2019) [21]. The model's performance is enhanced by optimizing a combined loss function that addresses the intricacies of both subtasks, effectively capturing the dual dependence of sentiment on both the aspect and target.

The model has five main parts: a BERT encoder that handles the input, two linear fully connected (FC) layers, and a softmax decoder that performs binary 'yes/no' classification. Additionally, there is an option to select either a conditional random field (CRF) decoder, following the approach of Ma and Hovy (2016) [22], or a softmax decoder, which is typically employed for sequence labeling tasks.

Furthermore, we broaden the scope of the TAS-BERT model to encompass a variety of model sizes, including those of BERT-Tiny, BERT-Medium, and BERT-Large [3]. In addition to leveraging the TAS-BERT model proposed by Wan and colleagues [17], we expand this methodology by employing various other pre-trained language models available in the Hugging Face Transformers library[4]. This modification involves utilizing Transformers models such as XLNet [23], ELECTRA [24], Sci-BERT [25], ConvBERT [26] and DistilBERT [27] in place of the BERT model. The integration of these diverse Transformers models has significantly broadened the development scope of the TAS approach. The architecture of the TAS-Transformers Model is adapted from the original architectures of TAS-BERT models proposed by Wan [17], with the inclusion of alternative pre-trained Transformers models from the Hugging Face

[3] https://github.com/google-research/bert
[4] https://github.com/huggingface/transformers

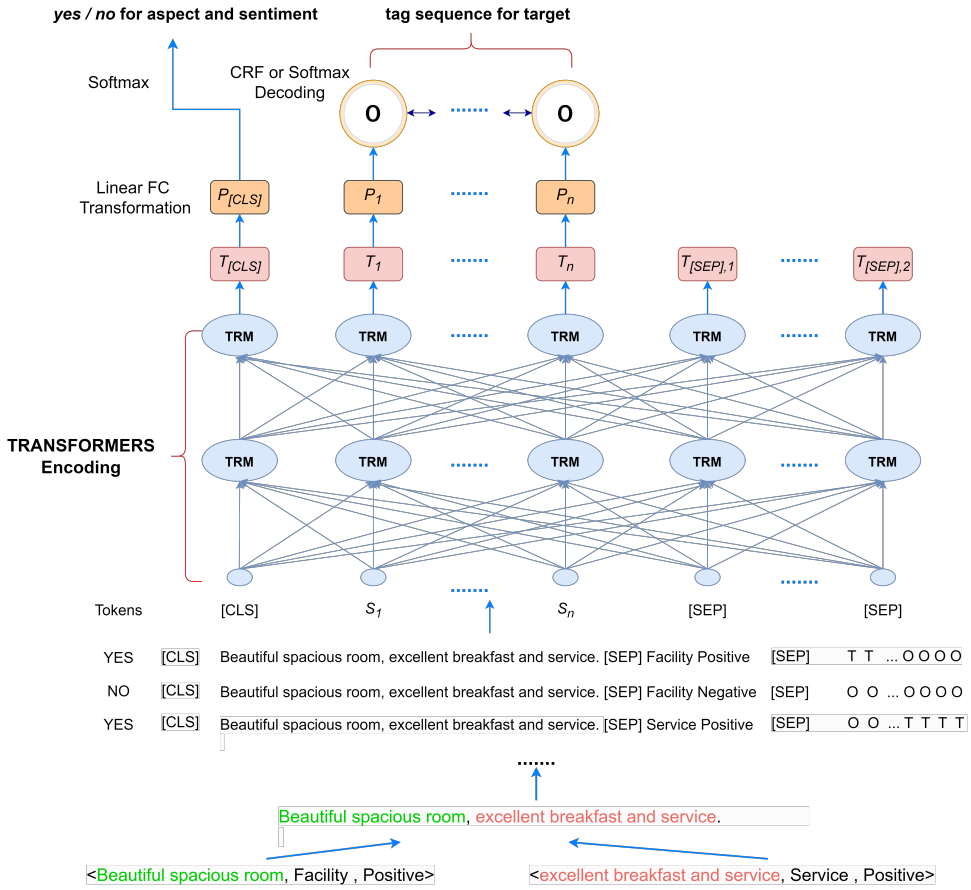

Fig. 2: The architecture and a running example for the TAS-Transformers model. TAS-Transformers takes a sentence-aspect category-sentiment polarity token sequence "[CLS]···[SEP]···[SEP]" as input. It outputs "yes/no" for predicting whether targets exist for the aspect-sentiment pair and a tag sequence for extracting the targets. [17]

Transformers library. This modified architecture is illustrated in Figure 2, enhancing flexibility and ease of method expansion.

4.2 Generative Aspect-based Sentiment analysis (GAS)

Target Aspect Sentiment Detection (TASD) [17] is a task that aims to identify and extract a triplet of (aspect term, aspect category, sentiment polarity) from a given input sentence.

In the paper of Zhang et al.,2021 [10], the authors propose two types of targets: annotation and extraction. While annotation is the process of labeling a triplet of aspect terms, aspect categories, and sentiment polarities in a sentence, extraction is the process of extracting the aforementioned triplets from that sentence. Herein, we

only use the extraction target due to its appropriateness since it allows the TASD model to be more flexible in identifying aspects and emotions in a sentence.

Given an input sentence x, GAS-Extraction predicts the complete set of aspect-level sentiment triplets, denoted as $T = \{AT, AC, SP\}$, such that AT, AC, SP is respectively aspect term, aspect category, and sentiment polarity. Among those three, an aspect term is a phrase that refers to a specific aspect of service or experience quality. Meanwhile, aspect category is a type of aspect, and sentiment polarity is the sentiment expressed toward that aspect. For example, in the sentence "The food at this restaurant is delicious," the aspect term is "the food at this restaurant," the aspect category is "quality," and the sentiment polarity is "positive".

This holistic approach captures all relevant aspects and sentiments within the sentence, providing a richer understanding of user opinions. The task is typically formulated as a supervised learning problem, in which the model is trained on a dataset of labeled sentences. The model then uses this training data to learn how to identify and extract the three components of the triplet from new, unlabeled sentences.

4.3 Aspect Sentiment Quad Prediction as Paraphrase Generation (ASQP)

ASQP [20] is a paraphrasing model that can identify the sentiment of a sentence about a specific aspect of a product or service. In our models, we inherit the idea of ASQP with two modifications. To begin with, instead of using the aspect term as a single word, which can be overly restrictive in certain cases, particularly when the aspect term comprises longer or more informative than a single word. Additionally, we define the aspect term as a contiguous subsequence of tokens in the sentence. This change allows the model to be more accurate in identifying the sentiment of a sentence.

Secondly, we modify the quadruplet representation from the original $Q \in \{c, at, ot, sp\}$ into $T = \{ac, at, sp\}$ in which ac, at, ot, sp respectively stand for aspect category, aspect term, opinion term, sentiment polarity. Specifically, we remove the opinion term from the quadruplet and introduce triplet. Given an input sentence x, the model aims to predict the set of all aspect-level sentiment triplets $T = \{ac, at, sp\}$ belonging to S_{ac}, S_x, S_{sp} respectively, where:

- $ac \in S_{ac}$ represents the aspect category, a discrete label from the predefined set of possible aspect categories.
- $at \in S_x$ represents the aspect term, a continuous span extracted from the sentence x. S_x denotes the set of all possible continuous spans in x, excluding the null case (no aspect mentioned)
- The sentiment polarity $sp \in \{POS, NEU, NEG\}$ is a class label that indicates the sentiment (positive, neutral, negative) expressed towards an aspect.

We use a variant of Transformers to leverage an encoder-decoder architecture to bridge the gap between natural language and a latent representation of its meaning. This representation captures rich semantic information. The pre-trained generation power of the model allows it to generate this type of language more accurately and efficiently:

$$\mathcal{L}_{ac}(ac) \text{ is } \mathcal{L}_{sp}(sp) \text{ because } \mathcal{L}_{at}(at) \text{ is } \mathcal{L}_{sp}(sp) \tag{1}$$

$\mathcal{L}_k$ converts each element of the triplet to a natural language form. $\mathcal{L}_k$ is a projection function of the element type k, which can be one of $\{ac, at, sp\}$. The specific conversion procedure is determined by a set of conditions. To ensure continuity between the aspect levels in a target sequence (y), the authors propose to include a special token [SSEP] between aspect-level sentiments. This token serves as a delimiter to indicate the boundary between different aspect levels.

So, our initial step involves transforming the task from the triplet format to natural language (y) for compatibility with the paraphrase framework employed by the model. To facilitate integration with the paraphrase framework, the linearization of each element within the sentiment triplet becomes necessary. This crucial step then essentially "flattens" the complex triplet structure, transforming it into a readily interpretable sequence of words. This enables the model to leverage its natural language processing capabilities for generating paraphrases as Table 2.

This process ultimately yields a set of transformed data points $D' = \{(x, y)\}$ where y represents the linearized, natural language representation of the sentiment initially encoded in the triplet.

Table 2: One example of the target sentence construction for the ASQP task [20]

Input-1	The room was very nice!
Label-1	(c, a, o, p) : (the room was very nice, facility, POS)
⇓	⇓
Target-1	facility is great because the room was very nice is great

For each aspect category (ac) and aspect term (at) in a triplet, we keep the natural language form. This is because we never encounter null values for these elements. For sentiment polarity (sp), we transform it as follows [20]:

$$\mathcal{L}_{sp}(sp) = \begin{cases} \text{great} & \text{if } sp = \text{POS} \\ ok & \text{if } sp = \text{NEU} \\ \text{bad} & \text{if } sp = \text{NEG} \end{cases} \tag{2}$$

We convert from class format to natural language representation to ensure that the target sequence (y) is logically and semantically coherent. This helps to reduce the likelihood that the sequence-to-sequence architecture will generate meaningless sentences.

Training model of both GAS-Extraction and ASQP framework: The Transformers architecture's [28] popularity for text-processing tasks stems from its two key components: the encoder and the decoder. Firstly, the encoder delves into the entire input's context, while, conversely, the decoder assembles the output sentence one token at a time.

Initiating the process, the encoder transforms the input sentence into a series of embeddings. These embeddings are then sequentially fed through self-attention

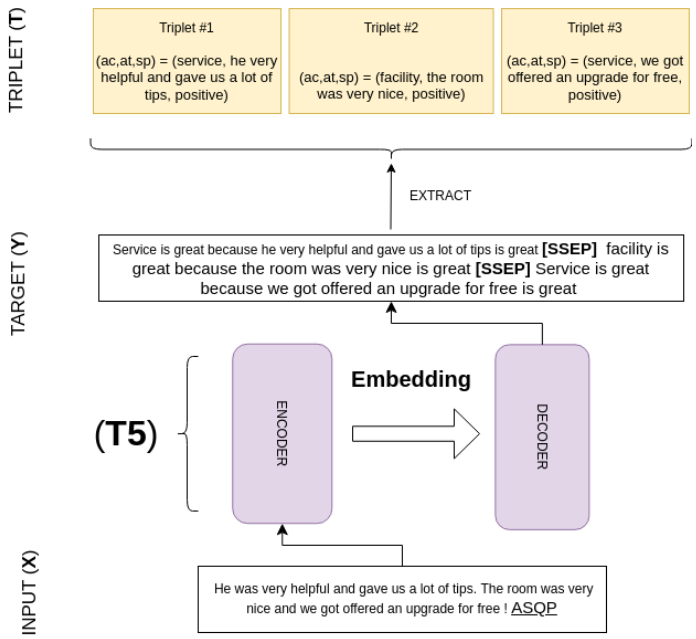

Fig. 3: Overview of the paraphrase generation framework (inspired from [20])

layers, enabling the encoder to grasp the intricate relationships between the input words. Subsequently, the decoder generates the output sentence by predicting the next token. This prediction relies on a conditional probability distribution, which carefully considers the previously generated tokens. The token with the highest vocabulary probability is then chosen for generation. Iteratively, this process continues until the complete sentence is formed.

Ultimately, the objective is to minimize the discrepancy between the generated and target sentences. This can be achieved using a loss function that quantifies the difference between the two. Finally, the model is updated via backpropagation to progressively reduce the loss.

Training phase: During the training process, we capitalize on the pre-trained knowledge of the T5 model [29], which has been trained on diverse datasets. We utilize its pre-trained weights as the initial weights (λ) for our model. Subsequently, we strive to fine-tune these weights optimally by fitting them to our specific input-target pairs. This fine-tuning process aims to maximize the log-likelihood function $p_\theta(\boldsymbol{y} \mid \boldsymbol{E}(\boldsymbol{x}))$, effectively guiding the model towards generating outputs that best align with the desired target sentences:

$$\max_{\theta} \log p_\theta(\boldsymbol{y} \mid \boldsymbol{E}(\boldsymbol{x})) = \sum_{i=1}^{l} \log p_\theta \left(\boldsymbol{y}_i \mid \boldsymbol{E}(\boldsymbol{x}), \boldsymbol{y}_{<i}\right) \tag{3}$$

where l is the length of target sequence y, and $E(x)$ is contextualized of input sequence (x).

Inference phase: During the inference process, we use the model with the optimal weight set to predict the test dataset. We then extract the triplet from the target sequence (y') and remove the special token [SSEP]. We then compare the extracted triplet with the gold sentiment T. If the extracted triplet does not meet the given conditions, it means that the generated sentence is not in the correct format. For example, if the sentiment polarity of the two clauses before and after the word "because" is not the same, or if one side of the sentiment polarity is generated but the other side is not, the triplet will be assigned as null. This ensures that the evaluation of the model is accurate and fair.

Evaluation: The authors evaluated the performance of ABSA models ([10], [20]) on the proposed dataset for generative models using the exact match metric. However, these methods are not suitable for our problem because the aspect term is a sequence of tokens with length $l > 3$. This means that the aspect term of two sentences may not be the same, even if they express the same sentiment. To address this issue, we use the sentence Bert technique [30]. This technique uses a pre-trained language model, called BERT, to compute the similarity between two sentences [31]. BERT first transforms

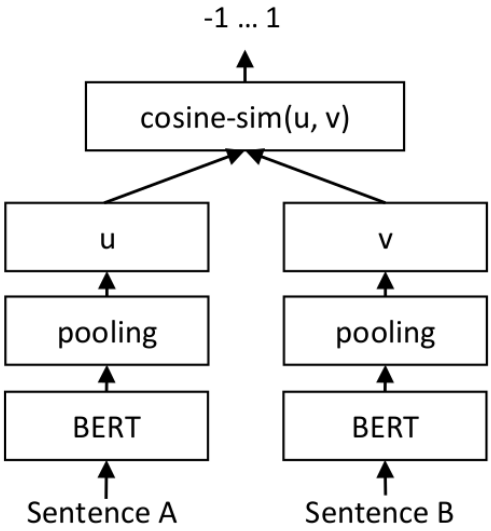

Fig. 4: Sentence Bert architecture [31]

each sentence into a contextualized encoded sequence (e). This sequence represents the meaning of the sentence in the context of the entire dataset. The similarity between the two sentences is then computed based on the cosine similarity between their contextualized encoded sequences. The formula for given two sentences a and b can be

defined as follows:

$$Sim(a,b) = Sim(e_a, e_b) = \frac{e_a.e_b}{||e_a||.||e_b||} > \alpha \tag{4}$$

We also introduce a lower bound α in Equation 4, which represents the minimum similarity threshold. If the cosine value is greater than α, we consider the two sentences to be similar. Otherwise, we consider them to be dissimilar.

While the sentence Bert technique is used to address the limitations of exact match for aspect terms, we maintain the use of exact match for both the aspect category and sentiment polarity in our evaluation.

5 Experiments

In this paper, most of our experiments were conducted on a computer equipped with an Intel Core i9 9900k processor with 64GB of RAM and an NVIDIA GeForce RTX 3090 GPU with 24GB of VRAM.

5.1 Dataset

We conduct experiments on a hotel domain dataset comprising 7,012 labeled sample reviews. For all methods, we split the dataset into a 70:30 ratio for training and testing, respectively. Preprocessing techniques are utilized to clean and normalize the formatted dataset, tailored to the requirements of each specific method. Table 3 describes the distribution of instances across sentiment polarity and aspect category.

Table 3: The distribution of instances across sentiment polarity and aspect categories.

	Training set				**Testing set**			
Aspect	Positive	Neutral	Negative	**Total**	Positive	Neutral	Negative	**Total**
Amenity	2,755	159	208	**3,122**	1,198	81	112	**1,391**
Branding	344	23	160	**527**	158	6	83	**247**
Experience	2,244	231	668	**3,143**	921	93	289	**1,303**
Facility	4,151	471	1,716	**6,338**	1,711	176	742	**2,629**
Loyalty	777	4	46	**827**	322	3	23	**348**
Service	5,950	307	942	**7,199**	2,548	134	385	**3,067**
Total	**16,221**	**1,195**	**3,740**	**21,156**	**6,858**	**493**	**1,634**	**8,985**

5.2 Results

We leveraged the following methods to address the TASD tasks and subtasks for our hotel reviews dataset. These approaches were meticulously selected and customized to effectively capture the diverse aspects of sentiment analysis present in our dataset.

- TAS-$BERT_{BASE}$ [17]: The TAS method, employing a BERT-based model for the ABSA problem, is another approach for our hotel reviews dataset. We leveraged the methodology outlined in the published code, using the Split Word (SW) method, the BIO tagging scheme, and a CRF decoder. This approach was employed to address the primary TASD task and encompass three subtasks: AD, ASD, and TSD.
- TAS-$BERT_{TINY}$, TAS-$BERT_{MEDIUM}$ and TAS-$BERT_{LARGE}$: Expanding the TAS method, we utilize various BERT sizes such as the $BERT_{TINY}$, $BERT_{MEDIUM}$, and $BERT_{LARGE}$ model to assess our dataset across ABSA tasks.
- TAS-XLNet: Instead of using the BERT model as previously described in the TAS method, we have adjusted the code of this method to utilize the XLNet base model available in the Hugging Face Transformers library.
- TAS-SciBERT: We have substituted a different Transformers model with a specific SciBERT uncased model from the Hugging Face Transformers library to evaluate its performance on the ABSA tasks.
- TAS-ELECTRA: Similar SciBERT, ELECTRA model from Hugging Face Transformers library is leveraged on this dataset.
- TAS-DistilBERT: We employed DistilBERT, a distilled version of BERT, to evaluate our dataset, benefiting from its efficient architecture without significantly compromising performance.
- TAS-ConvBERT: ConvBERT, a variant of BERT that incorporates convolutional neural networks, is utilized to assess our dataset, harnessing its unique architecture to potentially enhance performance in ABSA tasks.
- GAS-EXTRACTION [10]: We applied the published code from the GAS method to evaluate its performance on the hotel reviews dataset.
- ASQP: The modified ASQP method is employed in our experiments to assess its performance on both a new dataset and in a new domain, utilizing the published code.

Our experiments are done using various methods using the hotel reviews dataset. The performance of these approaches was compared based on the F1-score, which serves as the evaluation metric for all tasks on our hotel reviews dataset. The results of this comparison are detailed in Table 4. The output indicates that the TAS method demonstrates the best performance on our dataset, whether using the TAS-$BERT_{MEDIUM}$ model as described in the original paper or employing TAS with different transformers (excluding TAS-$BERT_{TINY}$). In contrast, two approaches, GAS-EXTRACTION and ASQP, exhibit lower performance on this dataset for the TASD task.

Results from the TAS-Transformer models: Leveraging various BERT model sizes and pre-trained language models available in the Hugging Face Transformers library in the TAS approach, along with a CRF decoder for sequence labeling and a BIO tagging scheme on the hotel reviews dataset, this method demonstrates superior performance on our dataset for the main TASD task and subtasks ASD, TSD, and AD. Comparison across Transformers models indicate that TAS-$BERT_{MEDIUM}$ achieves the best outcome on TASD and TSA tasks, while the TAS-$BERT_{BASE}$ model exhibits the highest output on the ASD and AD task. Moreover, there is not a significant disparity in the

Table 4: Comparison results of the TASD task across various methods. F1 scores(percent) are reported, with the best results highlighted in bold.

Method	Precision	Recall	F1-score
GAS-EXTRACTION	53.33	55.63	54.45
ASQP	55.99	58.63	57.27
TAS-$BERT_{TINY}$	65.33	41.73	50.93
TAS-$BERT_{MEDIUM}$	**89.39**	**76.39**	**79.74**
TAS-$BERT_{BASE}$	83.14	74.33	78.49
TAS-$BERT_{LARGE}$	77.36	64.96	70.62
TAS-XLNet	80.21	75.48	77.78
TAS-SciBERT	77.44	74.96	76.18
TAS-ELECTRA	78.56	73.54	75.96
TAS-DistilBERT	81.60	73.22	77.19
TAS-ConvBERT	78.90	74.96	76.88

results across all tasks among the Transformers models(excepting TAS-$BERT_{TINY}$ and TAS-$BERT_{LARGE}$) used in the TAS method on our dataset, as described in Table 5.

Table 5: Comparision results for four tasks on TAS-Transformer models. F1 scores (percent) are reported.

Method	TASD	ASD	TSD	AD
TAS-$BERT_{TINY}$	50.93	60.24	53.48	60.24
TAS-$BERT_{MEDIUM}$	**79.74**	79.76	**88.25**	79.76
TAS-$BERT_{BASE}$	78.49	**79.96**	85.69	**79.96**
TAS-$BERT_{LARGE}$	70.62	75.01	77.49	75.01
TAS-XLNet	77.78	77.81	87.72	77.81
TAS-SciBERT	76.18	76.21	87.63	76.21
TAS-ELECTRA	75.96	76.46	85.70	76.46
TAS-DistilBERT	77.19	77.35	85.84	77.35
TAS-ConvBERT	76.88	76.88	87.24	76.88

In addition, upon comparing the outcomes achieved by various methods on aspect categories in the hotel domain, differences in results among aspect categories become apparent. Generally, all methods exhibit the worst prediction output on the "Branding" aspect due to the limited number of samples in the dataset. Conversely, the "Service" aspect, which has the highest number of instances in the dataset, consistently outperforms other aspects for all approaches. Additionally, the two aspects, including "Facility" and "Amenity," also yield noteworthy results across all methods. Furthermore, each method demonstrates the best performance on specific aspects. GAS-EXTRACTION, ASQP, and TAS-$BERT_{TINY}$ methods illustrate outstanding results on aspect categories Service, Facility, and Amenity. Meanwhile, the remaining

methods exhibit highlighted output on Loyalty, Service, Facility, and Amenity. The results are depicted in Table 6.

The experiments clearly demonstrate the differences in prediction outcomes among methods regarding polarities. Across all approaches, Positive polarity consistently yields the best results, while Neutral sentiment ranks lowest in performance within our hotel reviews dataset. These outcomes are shown in Table 7.

Table 6: Classification report of TASD models on aspect categories. F1 scores are reported.

	Aspects								
Methods	Amenity	Branding	Experience	Facility	Loyalty	Service	Accuracy	Macro avg	Weighted avg
GAS-EXTRACTION	0.67	0.25	0.50	0.73	0.49	**0.75**	0.68	0.56	0.68
ASQP	0.68	0.26	0.53	0.74	0.57	**0.78**	0.71	0.59	0.70
TAS-$BERT_{TINY}$	0.66	0.00	0.20	0.58	0.00	**0.76**	0.60	0.37	0.56
TAS-$BERT_{MEDIUM}$	0.81	0.28	0.62	0.82	**0.88**	0.86	0.80	0.71	0.79
TAS-$BERT_{BASE}$	0.81	0.26	0.65	0.82	0.83	**0.86**	0.80	0.70	0.79
TAS-$BERT_{LARGE}$	0.78	0.00	0.52	0.77	**0.83**	0.82	0.75	0.62	0.74
TAS-XLNet	0.80	0.15	0.62	0.80	**0.87**	0.83	0.78	0.68	0.77
TAS-SciBERT	0.79	0.27	0.56	0.79	**0.87**	0.83	0.76	0.68	0.76
TAS-ELECTRA	0.79	0.12	0.59	0.78	**0.84**	0.83	0.76	0.66	0.76
DistilBERT	0.80	0.13	0.59	0.80	**0.87**	0.83	0.77	0.67	0.77
ConvBERT	0.79	0.30	0.57	0.79	**0.86**	0.84	0.77	0.69	0.76

Table 7: Classification report of TASD models on aspect polarity. F1 scores are reported.

	Polarity					
Methods	Positive	Neutral	Negative	Accuracy	Macro avg	Weighted avg
GAS-EXTRACTION	**0.94**	0.33	0.70	0.88	0.66	0.88
ASQP	**0.95**	0.32	0.71	0.89	0.68	0.89
TAS-$BERT_{TINY}$	**0.70**	0.00	0.00	0.60	0.24	0.54
TAS-$BERT_{MEDIUM}$	**0.85**	0.36	0.67	0.80	0.63	0.79
TAS-$BERT_{BASE}$	**0.85**	0.22	0.68	0.80	0.68	0.79
TAS-$BERT_{LARGE}$	**0.82**	0.00	0.53	0.75	0.45	0.72
TAS-XLNet	**0.83**	0.29	0.61	0.78	0.58	0.76
TAS-SciBERT	**0.82**	0.29	0.62	0.76	0.58	0.76
TAS-ELECTRA	**0.83**	0.23	0.59	0.76	0.55	0.75
DistilBERT	**0.83**	0.29	0.61	0.77	0.58	0.76
ConvBERT	**0.83**	0.31	0.61	0.77	0.58	0.75

Our hotel reviews dataset exhibits exceptional performance across various methods in ABSA. This highlights its potential as a valuable resource for diverse domains within ABSA. The dataset's robustness underscores its significance in not only benchmarking

existing approaches but also inspiring the development of innovative methods tailored to specific needs and applications

5.3 Error Analysis

To gain a deeper understanding of the method's performance in the context of our dataset, we conducted an error analysis in this section. This involved a careful examination of 100 samples from each model. Our primary objective was to identify and scrutinize the inaccuracies that emerged in task-specific outcomes.

5.3.1 Generative models

Our evaluation of several models revealed that generative models underperformed compared to only encoder models like BERT. Six key factors were identified as contributing to this performance gap:

- **Type I - Redundant Information Capture**: Instead of focusing on the relevant aspect level at time t, the model captures two aspect levels simultaneously—one at t and another at t+1. This redundancy can hinder effective analysis and prediction.
- **Type II - Error Amplification**: Due to the Type-I issue, the model's prediction at t+1 may be misled by its erroneous capture at t. Consequently, it attempts to extract information from t+2 instead of the intended aspect level at t+1, further propagating the error.
- **Type III - Temporal Misalignment**: Instead of focusing on the crucial aspect-level information at the current time step (t), it mistakenly extracts information from the previous time step (t-1). This offset creates a mismatch between the intended temporal context and the model's internal representation, leading to inaccurate predictions.
- **Type IV - Partial Information Loss**: Although the model identifies the target aspect segment, it loses key information during prediction. This incomplete representation leads to misclassification of the aspect category or sentiment.
- **Type V - Dissociation between Target and Features**: While the model captures the target aspect correctly, it misidentifies the associated aspect category or sentiment. This suggests a weakness in associating relevant features with the identified aspect.
- **Type VI - Aspect Branding Challenges**: Ambiguous or inconsistently labeled aspect levels pose a unique challenge for the model. This, coupled with potential limitations in its reasoning abilities, leaves certain cases uncovered and results in prediction errors.

In our prediction (Table 8), we found that five common types of errors accounted for the majority of errors. The sixth type of error, which was related to aspect branding, was relatively rare, occurring in only 21 cases.

5.3.2 TAS-Transformer models

The TAS-Transformer models exhibit a notable strength in accurately capturing targets. If the model correctly identifies the aspect-sentiment pair, it almost invariably

labels the target accurately, showing immunity to Types I, II, III, and IV errors. Therefore, the primary challenge for the TAS-Transformer models lies in making accurate predictions on aspect-sentiment pairs, which relates to Type V and especially Type VI errors.

In Table 9, it is observed that the models frequently struggle to accurately detect the aspect of branding when predicting the aspect-sentiment pair. This difficulty arises due to the similarity in meaning between the aspect of branding and those of facility and experience. Additionally, for the neutral sentiment, the TAS-Transformer models often become confused in choosing the correct sentiment. This confusion is particularly evident when the sentence contains the word "bad", leading the models to erroneously predict a negative sentiment. Similarly, the presence of the word "happy" in a sentence tends to incline the models toward predicting a positive sentiment.

6 Conclusion and Further Works

In this paper, we present a new dataset with a novel annotation scheme for the target. We believe that this dataset will provide a valuable resource for future research on the ABSA task. We have experimented with several methods for evaluating the quality of the dataset, and our results show that recent methods work quite well. However, we acknowledge that the dataset is relatively small, and it is possible that larger datasets could reveal new insights. Therefore, we plan to annotate more data in the future. In addition to these plans, we also intend to explore several other research directions. For example, we are interested in combining existing methods such as aspect-sentiment with target-aspect pairs. We believe that this combination could improve the performance of ABSA models. We are also aware of the limited number of research papers on the TASD task. To address this challenge, we plan to propose several new methods that are optimized not only for this dataset but also for previous datasets. Finally, we are excited about the potential of Large Language Models (LLMs) for ABSA. LLMs have been shown to be effective for a variety of natural language processing tasks, and we believe that they could also be effective for TASD. To explore this potential, we plan to experiment with LLMs to see whether they can improve the performance of ABSA models.

Declarations

Funding: There is no funding related to this article.
Conflicts of Interest: The authors have no competing interests to declare that are relevant to the content of this article.
Data Availability: The data related to this article can be provided upon request.

Table 8: Types of Errors in Generative Model. *AT*, **AC**, **SP** denote model incorrectly identified the aspect term, predicted the aspect category and predicted the sentiment polarity, respectively.

Text	Gold	Prediction	Type
very well .... no breakfast provided but nearby to restaurants ... nothing.	{*no breakfast provided*; service; **negative**}	{*no breakfast provided but nearby to restaurants*; service; **neutral**}	I
breakfast was great, the buffet was superior. the roof and outdoor pool are spectacular ... was old.	{*the buffet was superior*; **service**; positive}	{*the roof and outdoor pool are spectacular*; **facility**; positive}	II
staff service ... super friendly, accommodating, and organised. 10/10 would stay again because of the amazing staff!	{*10/10 would stay again*; **loyalty**; positive}	{*super friendly accommodating and organised*; **service**; positive}	III
the stay .... we were here for only one night so we didn 't really look around the whole place but it seemed like they had bbq and billiard tables ... if the name wasn 't misleading.	{*they had bbq and billiard tables*; **facility**; **positive**}	{*we were here for only one night so we didn't really look around the whole place but it seemed like they had billiard tables*; **experience**; **negative**}	IV
great stay ... we also enjoyed the roof top ... it was great!	{*we also enjoyed the roof top*; **experience**; positive}	{*we also enjoyed the roof top*; **facility**; positive}	V
one of our favourite places in Vietnam. .... stay there again!	{*one of our favourite places in Vietnam*; **branding**; positive}	{*one of our favourite places in Vietnam*; **experience**; positive}	VI

References

[1] Liu, B.: Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 1–167 (2012) https://doi.org/10.2200/s00416ed1v01y201204hlt016

[2] Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and

Table 9: Types of Errors in TAS-Transformer models. **AC**, **SP** denote model incorrectly predicted the aspect category, predicted the sentiment polarity.

Text	Gold	Prediction	Type
we didn't get the room that was in the photos of booking	{we didn't get the room that was in the photos of booking; **branding**; negative}	{we didn't get the room that was in the photos of booking; **experience**; negative}	VI
the hotel is not that new as it looks on the pictures	{the hotel is not that new as it looks on the pictures; **branding**; negative}	{the hotel is not that new as it looks on the pictures; **facility**; negative}	VI
facilities not too bad	{facilities not too bad; facility; **neutral**}	{facilities not too bad; facility; **negative**}	V
just simple hotel	{just simple hotel; facility; **neutral**}	{just simple hotel; facility; **positive**}	V
breakfast is fine if you're happy with coffee and fruit and yo	{breakfast is fine if you're happy with coffee and fruit and yo; service; **neutral**}	{breakfast is fine if you're happy with coffee and fruit and yo; service; **positive**}	V

Data Mining, pp. 168–177 (2004)

[3] Qiu, G., Liu, B., Bu, J., Chen, C.: Opinion word expansion and target extraction through double propagation. Computational Linguistics **37**(1), 9–27 (2011) https://doi.org/10.1162/coli_a_00034

[4] Liu, P., Joty, S., Meng, H.: Fine-grained opinion mining with recurrent neural networks and word embeddings. In: Màrquez, L., Callison-Burch, C., Su, J. (eds.) Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1433–1443. Association for Computational Linguistics, Lisbon, Portugal (2015). https://doi.org/10.18653/v1/D15-1168 . https://aclanthology.org/D15-1168

[5] Xu, H., Liu, B., Shu, L., Yu, P.S.: Double embeddings and CNN-based sequence labeling for aspect extraction. In: Gurevych, I., Miyao, Y. (eds.) Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 592–598. Association for Computational Linguistics, Melbourne, Australia (2018). https://doi.org/10.18653/v1/P18-2094 . https://aclanthology.org/P18-2094

[6] Zhou, X., Wan, X., Xiao, J.: Representation learning for aspect category detection in online reviews. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 29 (2015)

[7] Ruder, S., Ghaffari, P., Breslin, J.G.: A hierarchical model of reviews for aspect-based sentiment analysis. In: Su, J., Duh, K., Carreras, X. (eds.) Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 999–1005. Association for Computational Linguistics, Austin, Texas (2016). https://doi.org/10.18653/v1/D16-1103 . https://aclanthology.org/D16-1103

[8] Hu, M., Peng, Y., Huang, Z., Li, D., Lv, Y.: Open-domain targeted sentiment analysis via span-based extraction and classification. In: Korhonen, A., Traum, D., Màrquez, L. (eds.) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 537–546. Association for Computational Linguistics, Florence, Italy (2019). https://doi.org/10.18653/v1/P19-1051 . https://aclanthology.org/P19-1051

[9] Huang, B., Carley, K.: Parameterized convolutional neural networks for aspect level sentiment classification. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (eds.) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 1091–1096. Association for Computational Linguistics, Brussels, Belgium (2018). https://doi.org/10.18653/v1/D18-1136 . https://aclanthology.org/D18-1136

[10] Zhang, W., Li, X., Deng, Y., Bing, L., Lam, W.: Towards generative 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 2: Short Papers), pp. 504–510 (2021). https://aclanthology.org/2021.acl-short.64

[11] Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., Manandhar, S.: SemEval-2014 task 4: Aspect based sentiment analysis. In: Nakov, P., Zesch, T. (eds.) Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pp. 27–35. Association for Computational Linguistics, Dublin, Ireland (2014). https://doi.org/10.3115/v1/S14-2004 . https://aclanthology.org/S14-2004

[12] Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., Androutsopoulos, I.: SemEval-2015 task 12: Aspect based sentiment analysis. In: Nakov, P., Zesch, T., Cer, D., Jurgens, D. (eds.) Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pp. 486–495. Association for Computational Linguistics, Denver, Colorado (2015). https://doi.org/10.18653/v1/S15-2082 . https://aclanthology.org/S15-2082

[13] Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S.M., Eryiğit, G.: SemEval-2016 task 5: Aspect based sentiment analysis. In: Bethard, S., Carpuat, M., Cer, D., Jurgens, D., Nakov, P., Zesch, T. (eds.) Proceedings of the 10th International Workshop on Semantic

Evaluation (SemEval-2016), pp. 19–30. Association for Computational Linguistics, San Diego, California (2016). https://doi.org/10.18653/v1/S16-1002 . https://aclanthology.org/S16-1002

[14] Fan, Z., Wu, Z., Dai, X.-Y., Huang, S., Chen, J.: Target-oriented opinion words extraction with target-fused neural sequence labeling. In: Burstein, J., Doran, C., Solorio, T. (eds.) 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), pp. 2509–2518. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1259 . https://aclanthology.org/N19-1259

[15] Li, X., Bing, L., Li, P., Lam, W.: A unified model for opinion target extraction and target sentiment prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 6714–6721 (2019)

[16] Xu, L., Li, H., Lu, W., Bing, L.: Position-aware tagging for aspect sentiment triplet extraction. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2339–2349. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.emnlp-main.183 . https://aclanthology.org/2020.emnlp-main.183

[17] Wan, H., Yang, Y., Du, J., Liu, Y., Qi, K., Pan, J.Z.: Target-aspect-sentiment joint detection for aspect-based sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 9122–9129 (2020)

[18] Peng, H., Xu, L., Bing, L., Huang, F., Lu, W., Si, L.: Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 8600–8607 (2020)

[19] Zhao, H., Huang, L., Zhang, R., Lu, Q., Xue, H.: SpanMlt: A span-based multi-task learning framework for pair-wise aspect and opinion terms extraction. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 3239–3248. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.296 . https://aclanthology.org/2020.acl-main.296

[20] Zhang, W., Deng, Y., Li, X., Yuan, Y., Bing, L., Lam, W.: Aspect sentiment quad prediction as paraphrase generation. In: Moens, M.-F., Huang, X., Specia, L., Yih, S.W.-t. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 9209–9219. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (2021). https://doi.org/10.18653/v1/2021.emnlp-main.726 . https://aclanthology.org/2021.emnlp-main.726

[21] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran,

C., Solorio, T. (eds.) 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), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423 . https://aclanthology.org/N19-1423

[22] Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In: Erk, K., Smith, N.A. (eds.) Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1064–1074. Association for Computational Linguistics, Berlin, Germany (2016). https://doi.org/10.18653/v1/P16-1101 . https://aclanthology.org/P16-1101

[23] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems **32** (2019)

[24] Clark, K., Luong, M.-T., Le, Q.V., Manning, C.D.: ELECTRA: Pre-training text encoders as discriminators rather than generators. In: ICLR (2020). https://openreview.net/pdf?id=r1xMH1BtvB

[25] Beltagy, I., Lo, K., Cohan, A.: Scibert: A pretrained language model for scientific text. In: EMNLP. Association for Computational Linguistics, ??? (2019). https://www.aclweb.org/anthology/D19-1371

[26] Jiang, Z.-H., Yu, W., Zhou, D., Chen, Y., Feng, J., Yan, S.: Convbert: Improving bert with span-based dynamic convolution. Advances in Neural Information Processing Systems **33**, 12837–12848 (2020)

[27] Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In: NeurIPS EMC2 Workshop (2019)

[28] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[29] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21**(140), 1–67 (2020)

[30] Kenton, J.D.M.-W.C., Toutanova, L.K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT, vol. 1, p. 2 (2019)

[31] Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language

Processing (EMNLP-IJCNLP), pp. 3982–3992. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-1410 . https://aclanthology.org/D19-1410