Two New Large Corpora for Vietnamese Aspect-based Sentiment Analysis at Sentence Level

DANG VAN THIN, NGAN LUU-THUY NGUYEN, TRI MINH TRUONG, and LAC SI LE,
University of Information Technology, Vietnam National University, Ho Chi Minh City, VietNam
DUY TIN VO, Department of Computer Science, Lakehead University, Thunder Bay, ON P7B 5E1, Canada and VinAI Research, Ha Noi, VietNam

Aspect-based sentiment analysis has been studied in both research and industrial communities over recent years. For the low-resource languages, the standard benchmark corpora play an important role in the development of methods. In this article, we introduce two benchmark corpora with the largest sizes at sentence-level for two tasks: Aspect Category Detection and Aspect Polarity Classification in Vietnamese. Our corpora are annotated with high inter-annotator agreements for the restaurant and hotel domains. The release of our corpora would push forward the low-resource language processing community. In addition, we deploy and compare the effectiveness of supervised learning methods with a single and multi-task approach based on deep learning architectures. Experimental results on our corpora show that the multi-task approach based on BERT architecture outperforms the neural network architectures and the single approach. Our corpora and source code are published on this footnoted site.[1]

CCS Concepts: • **Computing methodologies** → **Language resources**;

Additional Key Words and Phrases: Aspect-based sentiment analysis, deep neural network, multi-task learning, Vietnamese corpora

ACM Reference format:
Dang Van Thin, Ngan Luu-Thuy Nguyen, Tri Minh Truong, Lac Si Le, and Duy Tin Vo. 2021. Two New Large Corpora for Vietnamese Aspect-based Sentiment Analysis at Sentence Level. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 20, 4, Article 62 (May 2021), 22 pages.
https://doi.org/10.1145/3446678

[1] https://sites.google.com/uit.edu.vn/uit-nlp/publications?authuser=0.

This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under grant number B2019-26-01.
Authors' addresses: D. V. Thin, N. L.-T. Nguyen (corresponding author), T. M. Truong, and L. S. Lac, University of Information Technology, Vietnam National University, Ho Chi Minh City, VietNam; emails: {thindv, ngannlt}@uit.edu.vn, {15520926, 17520669}@gm.uit.edu.vn; D. T. Vo, Department of Computer Science, Lakehead University, Thunder Bay, ON P7B 5E1, Canada and VinAI Research, Ha Noi, VietNam; email: tdvo@lakeheadu.ca.

2375-4699/2021/05-ART62 $15.00
https://doi.org/10.1145/3446678

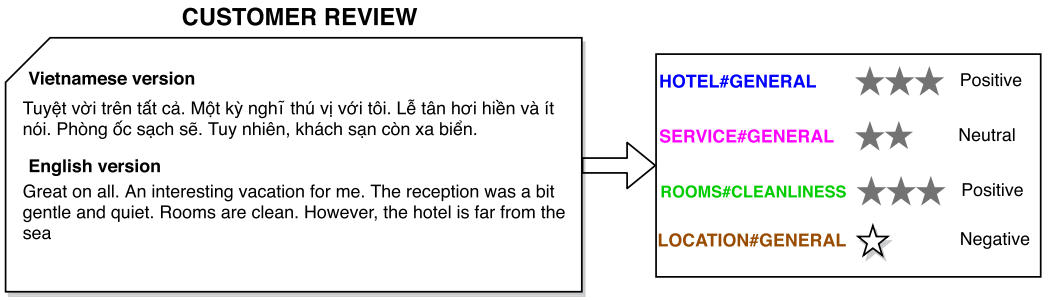

Fig. 1. An example review of the hotel domain in Vietnamese.

1 INTRODUCTION

With the rapid development of Information Technology, e-commerce sites and social media have been increasingly being exploited in many different ways to understand user insight in a popular way. Customers tend to express their comments and refer to other reviews related to a product or service online through those sites. In fact, these reviews are treated as valuable free resource and goodness measure for both organizations and consumers. We can explore the advantages and the disadvantages of products or services through practical experience of users from reviews. Each customer can comment on different opinions about the same product. Therefore, it is worth noting that comprehending user opinions is closely related to product improvement or customer campaigns of business organization (Ireland and Liu [8], Jin et al. [9]).

Aspect-based sentiment analysis *(ABSA)* (Liu and Zhang [14]; Pontiki et al. [26]; Pontiki et al. [27]) is an important task in sentiment analysis (known as opinion mining) to extract valuable information for providers and customers on users comment towards mentioning aspects [44]. It aims to identify the aspects of entities mentioned in a review and a sentiment polarity corresponding to these aspects in a certain domain. This allows us to perform a fine-grained analysis of review texts than traditional sentiment analysis. Most of text reviews are written as a short document that consist of many sentences. Therefore, it is practically useful to perform document-level aspect-based sentiment analysis, which predicts different sentiment for each aspect [47]. ABSA is divided into three main sub-tasks [26] as follows: Aspect Category Detection, Opinion Target Expression, Sentiment Polarity Classification. In this work, we only focus on Aspect Category Detection and Sentiment Polarity Classification sub-tasks in Vietnamese. We assume that the aspects have not been known, and our goal is to identify the aspect class mentioned reviews, then we assign the polarity for each detected aspect class. The number of aspect categories is predefined for each domain and the polarity consists of three labels: "positive," "negative," and "neutral." For example, in Figure 1, there are different aspects and corresponding polarities for a hotel review in Vietnamese.

The tasks in ABSA problem have attracted increasing interest in the NLP community these years. Most exhaustive studies have been done in English, but much fewer attempts have been made to tackle the aspect level task in other languages. We noted that the problem of data scarcity is significant, as any artificial intelligence project is based on data. The size of a dataset often causes poor performance in machine learning projects. Relevant data or the collection process, however, is too complicated and time-consuming. For that reason, we built two corpora at sentence-level related to the Vietnamese restaurant and hotel reviews to address our problem mentioned preceding. More specifically, we annotated near 10,000 sentences with high inter-annotator agreement for each domain for two sub-tasks in ABSA problem: Aspect Category Detection and Aspect Polarity Classification. To the best of our knowledge, this article is the first attempt to introduce

two large benchmark corpora at sentence-level for free research purpose in Vietnamese. In the natural language processing field, the multi-task approach has demonstrated the effectiveness by leveraging useful information among tasks. In this article, we want to explore the performance of multi-task approach based on the hypothesis that aspects contain relative information with others than the single approach. The single approach in our article means that each aspect will be trained by a separate architecture. We also experiment with supervised learning on two corpora with various basic deep neural architectures such as **Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM)**, and BERT model and its variants for evaluating our corpora. Our contributions of this work are summarized as follows:

- First, we present two new benchmark corpora for Vietnamese aspect-based sentiment analysis at sentence-level with good inter-annotator agreement on restaurant and hotel domains, respectively. To the best our knowledge, we are the first to build two sentence-level ABSA corpora in Vietnamese with the largest sizes (about 10,000 sentences).
- Second, we employ a supervised learning method to compare performance between the single and multi-task approach based on deep learning frameworks with inputs being Vietnamese domain-specific embeddings learned by our constructed data and BERT architecture that is trained on the large Vietnamese language.[2] From these results, we give the analysis of the effectiveness of different methods on this problem for Vietnamese.

The rest of this article is organized as follows: background and related work are introduced in Section 2. Section 3 presents the way to build two corpora, while Section 4 gives an overview of two approaches and detailed experiment. Results and discussion are provided in Section 5. Section 6 presents our discussion. Section 7 summarizes the article and provides our future research directions.

2 BACKGROUND AND RELATED WORKS

Aspect-based sentiment analysis (ABSA) has received much attention in the past decade. Many shared-tasks have been organized for ABSA with various tasks on different languages such as SemEval 2015 [27], SemEval 2016 [26], and Germeval 2017 [43]. Based on these shared tasks, ABSA is divided into three sub-tasks Aspect Category Detection, Opinion Target Expression, and Sentiment Polarity Classification. For general details, Aspect Category Detection is to identify pairs of an entity E and attribute to A (E#A) in a given text. E and A are chosen from predefined inventories of entity types (e.g., LAPTOP, MOUSE, RESTAURANT, FOOD) and attribute labels (e.g., DESIGN, PRICE, QUALITY). Opinion Target Expression is to extract the expression used towards each E#A pair. Sentiment Polarity last, term expression tuple, is to assign one of the following polarity labels: positive, negative, or neutral. The following example for restaurant domain will describe the output of each sub-task:

- **Input:** The pizza is very delicious but the staff is not friendly.
- **Output:**
  - —Aspect Category Detection: Food#Quality, Service#General
  - —Opinion Target Expression: pizza, staff
  - —Sentiment Polarity Classification: Food#Quality - positive, Service#General - negative, pizza - positive, staff - negative

These competitions also provided the variety of benchmark corpora corresponding to each subtask in broadly used languages such as English, Chinese, German, and so on. However, on less

[2] https://github.com/VinAIResearch/PhoBERT.

popular languages, there is a limitation in terms of corpora and their sizes. In early 2018, the VLSP shared-task was officially organized for ABSA in Vietnamese with two sub-tasks on two different domains [20]. The organizer built two benchmark document-level corpora with 4,751 reviews and 5,600 reviews for the restaurant and hotel domain, respectively. Their corpora are published for the community with free-research purpose.[3] Each corpus is annotated for two sub-tasks: aspect category detection and corresponding polarity classification task.

Several methods have been proposed to handle these competitions. For aspect category detection tasks, Zhou et al. [51] proposed a representation learning approach that automatically learns useful features from reviews on the SemEval-2014 dataset. In their paper, a semi-supervised word embedding is proposed to get continuous word representations. The authors used a logistic regression classifier combined with deeper and hybrid features based on neural networks to predict the aspect category. Yin et al. [47] transformed the ABSA tasks as a machine comprehension problem and proposed a hierarchical iterative attention architecture. Following that, Li et al. [11] also presented a novel HUARN architecture to jointly train the user preference and overall ratings. Their architecture is a type of hierarchy that can utilize the word-, sentence-, and document-level information. In addition, Wang et al. [41] applied the Hierarchical Reinforcement Learning approach for document-level Aspect Sentiment Classification task. This model can select the clause and word to remove the noise problem in the document-level review. Their experimental results demonstrated the potential of this approach for the ABSA problem. Wang et al. [40] presented a novel hierarchical architecture with both word-level and clause-level attentions by using sentence-level discourse segmentation technique. Then, Movahedi et al. [18] presented a deep attention mechanism neural network method to identify different aspect categories and outperformed existing methods on two benchmark restaurant datasets. Another interesting research work of Ghadery et al. [4] implemented a multilingual approach for aspect category detection in different languages by using Convolution Neural Network combined with the MUSE pre-trained word embedding. Recently, Chen et al. [2] proposed a **Cooperative Graph Attention Networks (CoGAN)** approach that incorporates two kinds of sentiment preference information for Aspect Sentiment Classification.

For the multi-task approach, there are some interesting research works proposed to solve these tasks. Xue et al. [45] proposed a multi-task learning MTNA model for aspect classification and aspect extraction. The MTNA defined aspect detection as a supervised classification and aspect extraction as a sequential labelling task and trained two tasks simultaneously based on a deep neural network. Their experiments demonstrated the effectiveness of their approach across three SemEval corpora. In addition, Li and Lam [13] presented a novel MIN framework—an LSTM-based deep multi-task learning model for aspect extraction. Their approach is based on the assumption that aspects and opinion word frequently always co-occur together. Therefore, they used the memory interaction between aspect term and opinion word extraction. Wang et al. [42] proposed the **Multi-task Memory Networks (MTNNs)** for aspect and opinion terms extraction. They developed an end-to-end deep learning architecture to exploit the commonalities and similarities of two tasks and achieved the new state-of-the-art performance on three benchmark corpora. Schmitt et al. [30] modeled the aspect detection and corresponding polarity in an end-to-end neural network. Additionally, they experimented with various neural network architectures (LSTM, CNN) and word representations (Word2vec, GloVe, fastText) on GermEval 2017 corpus. Their models achieved the new state-of-the-art results in GermEval corpus, showing that subword embedding (fastText) is crucial for morphological rich language. He et al. [5] explored the effectiveness of transfer knowledge by incorporating knowledge from the document-level corpus for aspect-based sentiment analysis. They demonstrated that the results could be improved when jointly training

[3] http://vlsp.org.vn/resources-vlsp2018.

document-level and aspect-level sentiment analysis. Recently, He et al. [6] proposed an **IMN (Interactive Multi-task learning Network)** architecture with message passing mechanism that can train many tasks at the same time based on informative interactions to send the useful information between tasks. This method showed good performance on three benchmark datasets. Yu et al. [48] designed a multi-task framework to extract the aspect terms and opinion terms simultaneously. They proposed the global optimization based on explicitly incorporating their intra-task and inter-task relations of two tasks. Moreover, Wan et al. [39] closely presented a novel method that trained three tasks in a joint architecture. This joint architecture is able to capture the dependence on both targets and aspects for sentiment polarity. Recently, there are many research studies showing the effectiveness of BERT model for ABSA problem (Sun et al. [32], Hoang et al. [7], Li et al. [12]).

For the low-resource language such as Vietnamese, there has been the little study of aspect-based sentiment analysis. Le et al. [10] proposed a semi-supervised method for aspect extraction and aspect classification based on the GK-LDA algorithm and dictionary. Their findings showed that their approach could detect aspects without depending on the length of the text and apply for other languages. We noticed that their model performs well when text documents do not contain too many slang words and clear meaning about specific aspects. However, a new perspective of using a sequence-labelling scheme associated with **bidirectional recurrent neural networks (BRNN)** and **conditional random field (CRF)** rather than rule-based approaches or conventional machine learning approaches with hand-designed features was proposed by Mai and Le [16]; it has turned over a new leaf in addressing the ABSA tasks. It requires no feature engineering efforts as well as linguistic resources, which allows them to adapt to other languages effortlessly. They used domain-specific language embeddings learned by using Word2vec on 2,098 sentences gathered from YouTube and pre-trained embeddings from Pham and Le-Hong [24] as an alternative. The drop-out technique is applied to embedding layers and all hidden layers before inputting into the other layers to avoid overfitting. They also did grid search with 5-fold cross-validation on the training set to tune other hyperparameters. Besides, Thin et al. [35] introduced a transformation method to participate in this shared-task and achieved the best scores for two sub-tasks of two domains. Their system consists of two components corresponding to each task where each component is composed of n binary classifiers with n the total number of aspects for the data domain. For the aspect detection task, the authors experimented with the **Support Vector Machine (SVM)** classifier with various handcrafted features (n-gram, word, part-of-speech) as a binary classifier. For the aspect polarity detection, they employed n multi-class classifier based on SVM with handcrafted features (n-gram, word feature, elongate word, aspect category, count of the hashtag, count of the POS feature, punctuation marks). However, their handcrafted features cannot capture the semantic information of polarity sentiment related to aspect. In addition, their approach does not take advantage of the relevant information between the aspects of the review. Then, Thin et al. [34] also proposed a deep CNN to solve the aspect detection task on two VLSP corpora. They modelled this task as a multi-label classification problem. The probability of final output is selected by a threshold to determine aspects for input text. Their experimental results show that deep learning methods can work well in these corpora. The limitation of their work is that the proposed model only solves the problem of detecting the aspects but cannot give the polarity sentiment of the detected aspect. In addition, Thuy et al. [36] presented a supervised method for aspect detection on their sentence-level annotated corpus. Their approach is similar to Thin et al. [35] but uses the different handcrafted features. Their method utilized annotated data from the English language and used word embedding features to capture the relationships between words. Besides, they also introduced a new annotated corpus for aspect detection in Vietnamese for restaurant domain with 3,796 sentences. Besides, other works of Nguyen et al. [21] also present an annotated corpus including 7,828 reviews at document-level with seven aspects combined with five polarity

Table 1. The Combination of the Entity and the Attribute for the Restaurant Domain

Acronym	Entity#Attribute	Acronym	Entity#Attribute	Acronym	Entity#Attribute
asp#1	Restaurant#General	asp#5	Food#Prices	asp#9	Drinks#Style&Option
asp#2	Restaurant#Prices	asp#6	Food#Style&Option	asp#10	Location#General
asp#3	Restaurant#Miscellaneous	asp#7	Drinks#Quality	asp#11	Ambience#General
asp#4	Food#Quality	asp#8	Drinks#Prices	asp#12	Service#General

sentiments for two tasks. They experimented an SVM method combined with handcrafted features on their corpus and achieved the F1-score 87.13% for aspect category detection and 59.20% for polarity sentiment analysis. In another study in this field, Tran and Phan [37] presented an ensemble learning model of sentiment classification with various features such as language features, sentiment shifting, and so on, for Vietnamese and English language.

From previous studies, we found that there is still a limited number of sentence-level corpora with large samples for ABSA in the Vietnamese language. Besides, the current approaches have not taken advantage of correlation information between the aspects, and no work has evaluated the effectiveness of the multi-task method between aspect detection and corresponding polarity at a sentence level in Vietnamese. For those reasons, in this article, we build two sentence-level corpora for two sub-tasks of ABSA for the restaurant and hotel domains. We deploy and experiment the single and multi-task model based on deep learning method inspired by Schmitt et al. [30] and Saeidi et al. [29] as our baselines. Besides, we also provide a detailed analysis of our corpora for further research.

3 CORPUS CONSTRUCTION

3.1 Data Collection

We collected the user feedback on popular websites about restaurants[4] and hotels[5] in Vietnamese. Especially, we focus on collecting reviews in major cities, because these cities have a large number of restaurants and hotels. These reviews are split into sentences by using UETSegmentation Library [22].[6] During the annotation process, we remove sentences that overlap with others that have been previously annotated and non-Vietnamese sentences. In addition, we exclude non-accent sentences. The following section describes the annotation guidelines and annotation process.

3.2 Annotation Schema and Guidelines

Our corpora are annotated by a group of workers and revised by linguistics experts. Each sentence will be assigned by two students and two others will revise and correct one more time. If there is an ambiguity between two annotations, then the final decision will be based on the discussion under the supervision of an expert. Given a sentence of the review, the annotators are required to assign the aspect category and aspect polarity labels. The number of aspects and polarity labels are chosen based on the shared-task SemEval 2016 [26] VLSP 2018 [20] competitions about Aspect-based sentiment Analysis task for many languages (e.g., English, Chinese, Russian) and Vietnamese language, respectively. For aspect category, there are 34 and 12 types of aspects (a combination of entity and attribute) corresponding to the hotel and restaurant domain. Table 1 and Table 2 show the aspect's names and its acronym for two domains. However, we decide to remove all sentences

[4] https://www.foody.vn/.
[5] https://mytour.vn/.
[6] https://github.com/phongnt570/UETsegmenter.

Table 2. The Combination of the Entity and the Attribute for the Hotel Domain

Acronym	Entity#Attribute	Acronym	Entity#Attribute	Acronym	Entity#Attribute
asp#1	Service#General	asp#13	Room_Amenities#Design&Features	asp#25	Facilities#Quality
asp#2	Rooms#Design&Features	asp#14	Hotel#Prices	asp#26	Facilities#Quality
asp#3	Rooms#Cleanliness	asp#15	Rooms#Comfort	asp#27	Ambience#General
asp#4	Hotel#General	asp#16	Rooms#General	asp#28	Food&Drinks#Miscellaneous
asp#5	Food&Drinks#Quality	asp#17	Hotel#Miscellaneous	asp#29	Room_Amenities#Cleanliness
asp#6	Facilities#General	asp#18	Hotel#Cleanliness	asp#30	Facilities#Cleanliness
asp#7	Hotel#Comfort	asp#19	Room_Amenities#Quality	asp#31	Facilities#Miscellaneous
asp#8	Hotel#Quality	asp#20	Hotel&Design&Features	asp#32	Food&Drinks#Prices
asp#9	Rooms#Quality	asp#21	Facilities#Design&Features	asp#33	Facilities#Prices
asp#10	Room_Amenities#General	asp#22	Food&Drinks#Style&Option	asp#34	Room_Amenities#Prices
asp#11	Location#General	asp#23	Rooms#Miscellaneous		
asp#12	Rooms#Prices	asp#24	Room_Amenities#Miscellaneous		

that express the user's emotions unrelated to the main object such as a restaurant or hotel. For the sentiment polarity, we select three types of polarity label: "Positive," "Neutral," and "Negative." Our guidelines are modified and combined based on the guidelines of the shared-task SemEval 2016 [26] and VLSP 2018 [20]. In the case of explicit reviews, the annotators directly apply the rules in the guidelines. However, for implied comments to describe as shown below, we need to revise and deeply analyze by a linguistics expert for the final decision.

- Theo mình phở ở những nơi khác bình dân mà ngon và rẻ vô cùng, chắc tại cái tiếng hồi xưa thôi chứ bây giờ thì tệ. *(In my opinion, phở in other places is quite cheap but delicious, probably because of the old reputation, but now it is bad.*)
- Nếu bạn nào muốn tiết kiệm, không tính đến loại khách sạn mấy sao có thể gửi oto tại bãi ngay trước khách sạn Ngọc Lan và liền kề đó có một loạt khách sạn mini cho các bạn lựa chọn, nếu không phải cuối tuần không cần đặt trước, phòng sạch sẽ, giá bằng phân nửa hay 1/3 giá Ngọc Lan. *(If you want to save money, regardless of the type of hotel, you can send a car at the beach right in front of Ngoc Lan hotel. There is a series of mini hotels for you to choose, if it was not on the weekend, you will not need a reservation, clean room, the price is half or 1/3 the price of Ngoc Lan.)*
- Nhân viên phục vụ rất chuyên nghiệp, đánh giá rất cao khả năng xử lý tình huống cũng như thái độ điềm tĩnh của các bạn ấy khi làm rơi gần như cái kéo (bị mất chốt nên bẻ làm 2) vào sát mặt mình mà vẫn rất bình tĩnh nhặt lên và không xin lỗi. (*The service staff is very professional, highly appreciating their ability to handle situations as well as their calm attitude when they drop almost the scissors (lost the pin, so they broke in 2) close to my faces but still very calmly picked up and did not apologize.*)

For the first sentence, user mentioned the *Food#Quality* and *Food#Prices* aspects with *Positive* polarity; however, this comment is comparing this restaurant's food with other restaurants. In case the restaurant is commenting, the review only mentions *Food#Quality* and has a *Negative* polarity. Similar to the previous sentence, the second review about the hotel domain also mention the good aspects of another hotel, but for the considered hotel, the aspect "Hotel#Prices" is assigned the "negative" sentiment. In the case of the third sentence, the review is meant to imply that the *Service#General* aspect is bad, but the user describes it scornfully complimenting service staff. Besides, our annotators face ambiguity in comments regarding the "Prices" attribute. We have

three price-related labels as Drinks#Prices, Food#Prices and Restaurant#Prices for the restaurant domain. In some cases of ambiguity, we have difficulty assigning the appropriate aspects. Looking at two following examples, we uniformly label it as {Food#Prices, neutral}, {Restaurant#Prices, negative} related to "Price" entity.

- Mình gọi tô phở tái, có giá là 40.000đ, mình cũng đã up menu cho các bạn, giá ở đây khá đắt so với những chỗ khác, nhưng mình ko đánh giá cao chất lượng. (I ordered phở, priced at 40,000 VND, I also uploaded the menu of the restaurant, the price here is quite expensive compared to other places, but I do not appreciate the quality.)
- Giá cả:,Mình thấy rẻ,gọi Top Blade(95k/150gr)là rẻ và ngon lắm rồi. (Price: I think it is quite cheap, I ordered Top Blade (95k/ 150gr), which was cheap and delicious.)

Another problem we encountered in the annotating process was that the data was at the sentence level and separated from the document-level comments. Therefore, our annotators will infer to select proper labels for aspects in the considering sentence by using their experience (see the following examples; each domain has two examples). In the process of annotating, we will automatically remove the comments that are not related to the restaurant object (users' reasons, etc.) from our corpora.

- Bỏ cả dĩa, khô và dai lắm. (Discard the plate, dry and chewy.) is annotated by the {Food#Quality, negative} label.
- Quá ngon và quá tuyệt vời đúng không nào? (Too delicious and too great, right?) is annotated by the {Restaurant#General, positive} label.
- Tôi chấm cho dịch vụ của khách sạn Ngọc Lan Đà Lạt 9 điểm để cho khách sạn có thể phấn đấu chứ thực tế thì chẳng có điểm nào mà tôi cảm thấy chưa hài lòng cả. (I gave the service of Ngoc Lan Hotel to Dalat 9 points so that the hotel could strive for it. In fact, there is no point that I feel is not satisfied at all.) is assigned by {Hotel#General, positive}.
- Tôi cảm thấy thật vui khi đã chọn được 1 dịch vụ tốt như vậy (I feel so happy to have chosen such a good service) is also assigned the label {Hotel#General, positive}.

The annotators will go through three stages (stage 1, stage 2, stage 3) of training annotators and guidelines will be updated based on our discussion. If there is ambiguity between annotators, then the final decision will be decided by experienced labelling experts. Through each stage, Cohen's kappa coefficient [1], which is a criterion to measure the quality of labelling, is calculated the inter-annotator agreement on a sample set and discuss to solve ambiguous cases in the annotating process. The guidelines will also be updated based on the author's agreement and annotators. The Cohen's kappa coefficient [1] is calculated as follows:

$$A_m = \frac{P_o - P_e}{1 - P_e}. \tag{1}$$

Pairwise inter-annotator agreement for two domains is shown in Table 3. A_m is the inter-annotator agreement, P_o is the observed agreement between annotators, and P_e is expected agreement. As shown in Table 3, we achieved a high inter-annotator agreement for two tasks of two domains. Those values demonstrate that our corpora are reliable and eligible to be two new standard benchmark corpora for ABSA in Vietnamese.

3.3 Corpora Statistics

In this section, we present the statistics of the annotated data. In total, we annotated 10,005 and 9,737 sentences for the hotel and restaurant domain, respectively.

Table 3. The Inter-annotator Agreements on Our Corpora for Two Domains (in %)

Domain	Task	P_o	P_e	A_m
Restaurant	Aspect Category	95.26	65.35	86.32
	Aspect Polarity	97.15	87.26	77.63
Hotel	Aspect Category	96.36	74.85	85.53
	Aspect Polarity	93.52	63.97	82.01

Table 4. The Distribution of the Aspects in Our Corpora for Two Domains

The Information	Restaurant			Hotel		
	Train	Dev	Test	Train	Dev	Test
No. Reviews	7,028	771	1,938	7,180	795	2,030
No. Token	115,753	12,669	31,784	131,020	14,739	37,083
No. Vocab	9,260	2,834	4,679	4,034	1,570	2,437
No. Aspects	9,458	1,053	2,629	11,812	1,318	3,283
Avg. Aspect per Sentence	1.35	1.48	1.36	1.65	1.66	1.62
Average Length	16.47	17.82	16.40	18.25	18.54	18.27

Table 5. The Distribution of Aspect and Its Polarity in the Training, Development, and Test Set for Restaurant Domain

Aspect	Train			Development			Test			Total
	Positive	Neutral	Negative	Positive	Neutral	Negative	Positive	Neutral	Negative	
AMBIENCE#GENERAL	534	115	144	72	24	17	148	32	47	**1,133**
DRINKS#PRICES	30	102	33	6	14	4	9	29	9	**236**
DRINKS#QUALITY	462	131	116	71	20	10	122	42	39	**1,013**
DRINKS#STYLE&OPTIONS	171	238	43	26	34	5	53	62	14	**646**
FOOD#PRICES	89	245	56	15	37	4	25	76	11	**558**
FOOD#QUALITY	1,343	316	281	196	36	45	390	69	95	**2,771**
FOOD#STYLE&OPTIONS	721	597	210	105	80	33	222	148	67	**2,183**
LOCATION#GENERAL	173	159	31	21	23	8	43	54	7	**519**
RESTAURANT#GENERAL	499	238	140	78	29	18	165	59	27	**1,253**
RESTAURANT#MISCELLANEOUS	286	76	146	44	11	18	80	21	44	**726**
RESTAURANT#PRICES	130	186	93	24	22	13	38	52	27	**585**
SERVICE#GENERAL	608	88	366	91	9	52	171	28	104	**1,517**
Total	**5,046**	**2,491**	**1,659**	**749**	**339**	**227**	**1,466**	**672**	**491**	

We need a strategy to divide our corpus into three sub-sets with a uniform distribution over aspects. To address this problem, we use an iterative algorithm[7] for the stratification in the multi-label corpus that is proposed by Sechidis et al. [31]. Our corpora are divided into training, development, and test sets with the ratio of 7/1/2. Table 4 shows the basic statistics of the two datasets. Table 5 and Table 6 present the distribution of aspects with its polarity in the training, development, and test datasets. As shown in Table 5 and Table 6, it is easy to see that the number

[7]https://github.com/trent-b/iterative-stratification.

Table 6. The Distribution of Aspect and Its Polarity in the Training, Development, and Test Set for Hotel Domain

Aspect	Train			Development			Test			Total
	Positive	Neutral	Negative	Positive	Neutral	Negative	Positive	Neutral	Negative	
SERVICE#GENERAL	1,906	129	266	221	16	20	524	31	83	**3,196**
HOTEL#GENERAL	932	219	62	109	23	4	259	60	17	**1,685**
ROOMS#CLEANLINESS	931	6	130	100	0	17	253	2	41	**1,480**
LOCATION#GENERAL	843	19	95	90	3	14	230	4	31	**1,329**
HOTEL#COMFORT	811	15	93	94	0	8	218	12	26	**1,277**
ROOM_AMENITIES#GENERAL	651	17	36	70	2	6	188	2	5	**977**
HOTEL#PRICES	557	27	46	61	3	6	149	12	14	**875**
ROOMS#DESIGN&FEATURES	328	8	151	35	2	17	86	6	43	**676**
ROOMS#COMFORT	275	2	111	26	1	16	72	0	35	**538**
FOOD&DRINKS#QUALITY	241	36	89	28	2	11	65	10	27	**509**
HOTEL#QUALITY	194	37	54	20	7	6	55	11	13	**397**
HOTEL#DESIGN&FEATURES	258	5	22	28	0	4	65	2	12	**396**
FOOD&DRINKS#STYLE&OPTIONS	109	12	146	12	1	17	27	2	46	**372**
ROOMS#GENERAL	172	44	31	17	4	6	50	11	6	**341**
ROOM_AMENITIES#QUALITY	67	7	146	11	0	15	17	1	44	**308**
ROOMS#QUALITY	147	13	53	14	3	7	42	2	16	**297**
HOTEL#CLEANLINESS	164	7	16	16	2	4	46	2	4	**261**
FACILITIES#GENERAL	128	10	23	15	2	2	37	3	6	**226**
OTHER ASPECTS	5,876	265	1,242	637	32	156	1,600	82	369	**10,259**
Total	**8,714**	**613**	**1,570**	**967**	**71**	**180**	**2,383**	**173**	**469**	

"OTHER ASPECTS" denotes the total number of aspects whose frequency appears less than 200 times in the whole corpus.

of aspects with polarity labels is evenly divided among different training, development, and test datasets. In terms of numbers of samples per aspect, we can observe that the iterative stratification strategy evenly distributes reviews for three datasets, which allows the learning algorithms to be evaluated objectively. However, considering the distribution of aspects in the whole corpus, we see that there is an imbalance between them. For example, the Drinks#Prices and Facilities#General aspects have few reviews in comparison with the other aspects for the restaurant and hotel domain. This is an acceptable phenomenon of corpus creation, however, it reveals a challenge of our corpora.

In addition, we also analyzed the review length (in the sentence) and the number of aspects in a review. As shown in Figure 2, the numbers of reviews that contain only one aspect dominate the two corpora (69% and 60% for the restaurant and hotel domain, respectively). We found that users tend to express their opinions about a restaurant or a hotel with a paragraph and each sentence often refers to only one or two different aspects. The percentage of sentences with one or two labels makes up the majority of our corpora at sentence level. The "other" value indicates the proportion of the review that is assigned with more than three aspects. In terms of the review length, sentences with length from 10 to 20 tokens occupy the highest proportion with 45% and 47% for the restaurant domain and the hotel domain, separately.

We are also interested in the challenging level of our corpora. We found that reviews that contain at least two aspects with different polarities account for 14.56% and 6.07% of the restaurant and hotel corpora, respectively. Such reviews are challenging to the models. They require that the model correctly identify different sentiment polarities of different aspects in the same sentence.

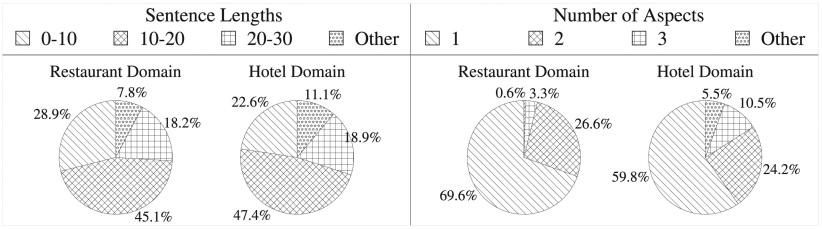

Fig. 2. Charts give information on the percentage of each aspect and the percentage of the length of the review in the whole corpus. The left and the right box are the charts of the restaurant and hotel domain, respectively.

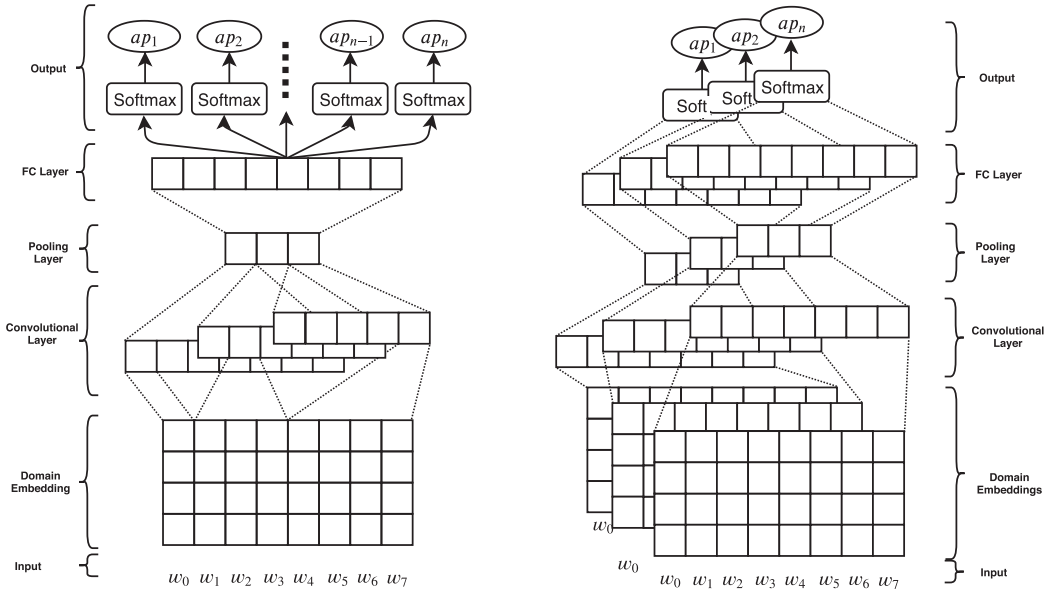

Fig. 3. Two approaches to solving the ABSA tasks in our corpora. The model on the left is CNN architecture based on the multi-task approach to predict the aspect and its polarity. The remaining model is a combination of CNN architectures based on a single approach to predict each aspect and corresponding polarity.

4 METHODS

To provide baseline results on our datasets, we performed experiments to evaluate different machine learning methods on our annotated corpora including a supervised method, different single, and multi-task deep learning models. Given an input sentence, the model outputs M aspect categories as well as the corresponding polarities of N polarities. In the single model approach, we constructed M classifiers with M sets of parameters. Each classifier has (N+1) outputs that determine whether the input sentence contains a particular aspect and corresponding polarity. In the multi-task model approach, one shared-parameter model is employed with M sets of (N+1) outputs corresponding to M x (N+1) output units to encode aspect categories and polarities.

Figure 3 shows the architectures of two approaches to solve two tasks. In detail, we selected the following methods for comparison.

- **Multiple SVM:** Thuy et al. [36] and Thin et al. [35] presented a transformation method based on multiple binary classification. This approach uses SVM classifiers with various

handcrafted features: n-gram, word, **part-of-speech (POS)** information for aspect detection and n-gram, word, elongate word, aspect category, count of the hashtag, count the POS feature, punctuation mark for sentiment polarity. In this article, we re-implemented their approach for the aspect detection and corresponding polarity task.

- **Single LSTM:** Saeidi et al. [29] presented a novel approach based on LSTM architecture that can predict the aspect and corresponding polarity. For each aspect, the authors introduced a new polarity class "None" to indicate that the aspect is assigned to input.
- **LSTM:** Schmitt et al. [30] showed the effectiveness of multi-task approach using Long Short-Term Memory with the fastText pre-trained embedding.[8] Their results showed that subword information is crucial for word representation in a morphologically rich language. For the Vietnamese language, we collected the user reviews and trained the domain-specific word embedding using Skip-gram method for all experiments.
- **CNN:** Schmitt et al. [30] also presented a different version of end-to-end model based on Convolutional Neural Network (CNN) that can be trained for aspect detection and polarity detection simultaneously.

For representing the output of two sub-tasks, inspired by Schmitt et al. [30] and Saeidi et al. [29], we conducted a series of different architectural models based on the deep neural network as follows:

- **BiLSTM + Attention:** We implemented a minimalistic model based on Bidirectional **Long short-term Memory (BiLSTM)** combined with an attention (Yang et al. [46]). The inputs are fed into bidirectional LSTM to acquire the context. After that, we combined the vector representation of attention layer, max pooling, and mean pooling to produce the input representation before putting them in the fully connected layer.
- **LSTM + Attention:** Similar to the BiLSTM with Attention architecture, we replaced the BiLSTM layer with LSTM to memorize the context of the input.
- **CNN-LSTM:** Zhou et al. [50] presented an unified architecture based on CNN and LSTM to capture both local features of phrase and global sentence semantics. In this article, we also implemented this architecture to explore its effectiveness for the two tasks.
- **CNN-LSTM-Attention:** Another variant of CNN-LSTM architecture is CNN-LSTM combined with the attention on the top layer of LSTM. The attention layer closely follows the Yang et al. [46] work.
- **BiLSTM - CNN:** This is a popular model based on two mainstream architectures, CNN and RNN. We used BiLSTM to extract the contextual information of sentences, then fed them into the multi-channel CNN layer to get the local features. In addition, we applied global max pooling concatenate with global average pooling to represent the input.
- **BERT:** Recently, BERT architecture has shown the effectiveness for various down-stream tasks for Natural Language Processing. In this article, we use the PhoBert,[9] [19] which is trained for Vietnamese language. We extracted the hidden state of [CLS] token as representation of the whole sentence and put them into linear layer with softmax activation.

Experimental Settings: All architectures are installed according to the description mentioned in the corresponding papers. All models share the same data pre-processing component; specifically, we applied the steps as presented by Thin et al. [35]. For the CNN architecture, we used the filter windows of 3, 4, 5; each filter has 128 feature maps. For activation, we set the ReLU activation as the nonlinear function in convolution layers and a fully connected layer. For LSTM

[8] https://fasttext.cc/docs/en/crawl-vectors.html.
[9] https://github.com/VinAIResearch/PhoBERT.

Table 7. The Results of Aspect Detection Task Based on the Single and Multi-task Approach for Restaurant Domain (in %)

Models	Multi-task Approach			Single Approach		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Multiple SVM	-	-	-	76.14	77.44	76.79
LSTM	84.49	80.28	82.33	82.87	76.62	79.61
CNN	83.92	79.40	81.59	83.21	78.73	80.91
LSTM + Attention	83.67	81.53	82.58	83.90	77.32	80.48
BiLSTM + Attention	83.16	83.31	83.23	83.76	77.27	80.39
CNN-LSTM	85.43	79.40	82.31	84.18	78.82	81.41
CNN-LSTM + Attention	85.27	80.64	82.89	83.57	81.46	82.50
BiLSTM-CNN	86.06	81.48	**83.70**	85.22	78.63	81.80
BERT	89.17	84.86	**86.96**	87.41	82.16	84.71

Table 8. The Results of the Aspect with Corresponding Polarity Sentiment Task for Restaurant Domain (in %)

Models	Multi-task Approach			Single Approach		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Multiple SVM	-	-	-	59.35	60.37	59.85
LSTM	64.74	61.52	63.09	61.22	56.60	58.81
CNN	68.61	64.91	66.71	67.34	63.71	65.48
LSTM + Attention	68.26	66.51	67.37	67.34	62.06	64.59
BiLSTM + Attention	66.74	66.86	66.79	66.97	61.78	64.27
CNN-LSTM	65.94	61.29	63.53	65.04	60.90	62.90
CNN-LSTM + Attention	67.49	63.83	65.61	66.12	64.82	65.46
BiLSTM-CNN	71.47	67.67	**69.52**	69.63	64.25	66.83
BERT	76.78	73.07	**74.88**	74.18	69.72	71.88

architecture, the recurrent units are set to 256. The embedding and fully connected layers use dropout technique with the probability of 0.5 and 0.2, respectively. The batch sizes and the number of epochs are set to 50 and 100, respectively. Comparing with previous works, we use the new state-of-the-art **Rectified Adam (RAdam)** optimizer for deep neural network framework that is proposed by Liu et al. [15]. For word representation, we collected a large number of domain reviews and trained the word embedding using Řehůřek and Sojka [28] framework. Our training data include 227,995 pre-processed sentences with 3,491,716 tokens and 3,388,046 pre-processed sentences with 55,636,954 tokens corresponding to the hotel and restaurant domains, respectively. In all experiments, word representation was trained by Skip-gram [17] architecture with default parameters (100-dimensional vector, 5-window size). For the BERT architecture, we use the BERT-base with 12 Transformer blocks with 12 self-attention heads, and hidden size of 768.

5 RESULT AND ANALYSIS

5.1 Experimental Results

In this section, we present the experimental results for the aspect detection task and the aspect polarity task. The results for the restaurant domain are shown in Table 7 and Table 8. Meanwhile, Table 9 and Table 10 present the results of two tasks for the hotel domain.

Table 9. The Results of Aspect Detection Task Based on the Single and Multi-task Approach for Hotel Domain (in %)

Models	Multi-task Approach			Single Approach		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Multiple SVM	-	-	-	76.68	74.70	75.68
CNN	78.61	74.35	76.42	80.19	72.94	76.39
LSTM + Attention	83.47	69.07	75.59	79.12	71.65	75.20
BiLSTM + Attention	82.02	72.08	76.73	80.91	72.36	76.40
CNN-LSTM	10.74	42.35	17.14	84.37	22.83	35.94
CNN-LSTM + Attention	76.92	70.76	73.71	77.07	76.56	76.81
BiLSTM-CNN	77.11	78.22	**77.66**	81.03	73.15	76.89
BERT	83.46	75.18	79.10	79.02	76.31	77.64

Table 10. The Results of the Aspect with Corresponding Polarity Sentiment Task for Hotel Domain (in %)

Models	Multi-task Approach			Single Approach		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Multiple SVM	-	-	-	69.06	67.28	68.16
CNN	71.48	67.61	69.49	72.57	66.01	69.14
LSTM + Attention	76.22	63.07	69.03	72.52	65.68	68.93
BiLSTM + Attention	74.68	65.63	69.86	73.82	66.01	69.70
CNN-LSTM	07.72	30.43	12.32	77.01	20.84	32.80
CNN-LSTM + Attention	69.02	63.50	66.14	69.27	68.80	69.03
BiLSTM-CNN	70.23	71.23	**70.72**	73.01	65.92	69.29
BERT	77.75	70.03	73.69	72.20	69.72	70.94

We can see that the multi-task approach is better than the single approach for most of the models. Based on the basic CNN architecture, the multi-task approach achieved better F1-score than the single approach for both of the tasks and both of the domains. For the aspect detection, the multi-task approach based on the CNN model brings the relative improvements of 0.68% and 0.03% in term of F1-score than the single approach based on the CNN model for the restaurant and hotel domain, respectively. Compared with the single approach, the multi-task approach also gives the improvements of 1.23% and 0.35% for aspect with polarity detection task. Except for the CNN-LSTM architecture on the hotel domain, this architecture is not effective for single and multi-task approach. Specifically, the result of CNN-LSTM model based on the single approach only can predict two labels including Service#General and Hotel#Comfort, because these have been mostly assigned in the training set. Another reason is that sentence patterns assigned with this label have quite a similarity. Therefore, CNN-LSTM architecture can only predict sentences belonging to the above two aspects. Meanwhile, the multi-task approach can predict evenly between aspects but the results are not high. It is, moreover, amazing that the baseline LSTM architecture for both approaches cannot correctly assign any aspect for the hotel domain. Overall, it proves that the deep neural architecture based on the multi-task approach is more effective for our corpora than the single approach.

Compared with the supervised learning method with handcrafted features, most of the neural network architectures achieved impressive results on the two corpora. This shows the potential of applying the neural network models to our corpora. Among all the mentioned neural network architectures in Section 4, the BiLSTM-CNN is the best baseline method for our corpora amongst the

multi-task models. As shown in Tables 7 and 9, this architecture based on multi-task approach is better than the others in improving F1-score from 0.47% to 2.11% and 1.24% to 3.95% for the restaurant and hotel domain, respectively. Similarly, BiLSTM-CNN also boosts the F1-score from 0.94% to 5.22% and 0.86% to 4.58% more than other neural models, in Table 8 and Table 10. Comparing neural network models with BERT architecture, we can see that the BERT model demonstrates its effectiveness in ABSA problem. In detail, BERT model improves the F1-score of +3.26% and +5.36% for the aspect category detection task in multi-task approach than BiLSTM-CNN model. For the aspect with corresponding sentiment task, BERT model also is higher than the BiLSTM-CNN model with the improvement of F1-score (+1.44% and +2.97% for the restaurant and hotel domain, respectively). This demonstrated the effectiveness of BERT architecture in the ABSA problem.

5.2 Analysis

Looking at the results, it can be seen that the attention layer has a powerful influence on the performance of LSTM and BiLSTM architecture in the two approaches. For the restaurant domain, we observed that the performance of LSTM increases at least 0.25% for aspect category detection task and at least 5.78% for the aspect category and corresponding polarity task. For the hotel domain, the LSTM architecture cannot predict correctly any input sentence; however, when we use the output of LSTM as the input attention layer, the results improve significantly. In detail, the multi-task approach based on CNN-LSTM architecture [50] without attention layer just achieves the F1-score of 17.14% and 12.32%, while the single approach achieves the F1-score of 35.94% and 32.80% for the two tasks, respectively. However, adding the attention layer after the output of the LSTM layer helps to increase the performance of original architecture. Similar to the LSTM architecture, attention layer also helps the BiLSTM model increase the performance of the two approaches. It proves that the attention layer is useful for LSTM and BiLSTM architecture in our corpora.

In addition to evaluating the overall results on the testing set, we also investigated the effectiveness of the best multi-task approach based on the BERT architecture according to the data analysis cases in Section 3. First, we analyzed the performance of best method related with respect to the sentence length of the sentences and the number of aspects mentioned. Second, we explored the effectiveness of the best method on the subsets of the corpora that contain sentences with at least two aspects with different polarities. All analyses were carried out entirely on the development set.

The distribution according to the length of the sentences corresponding to the restaurant and hotel domain as follows: 29.57% and 22.24% for range 1 to 10 words, 43.84% and 47.55% for range 10 to 20 words, 18.55% and 19.37% for range 20 to 30 words, and 8.04% and 10.81% for the remaining cases. For the distribution of the aspect, two domains have the corresponding values as follows: 67.57% and 58.11% for one aspect, 28.40% and 25.91% for two aspects, 3.89% and 10.19% for three aspects, 0.12% and 5.78% for remaining cases. Figures 4 and 5 show the results of two cases for two domains. From Figure 4, it is clear that the performance decreases according to the sentence length. However, for the restaurant domain, it is surprising that the performance on sentences with length larger than 30 tokens outperforms those of the sentence lengths 10–20 and 20–30. We found that most aspects of this subset have a high frequency in the training set. Contrary to the results of the sentence length, the results of the total number of aspects for each sentence give good results for the sentences with more than one aspect. This can be partly explained that each sentence is often labelled with two and three aspects that have the same entity, such as Food#Quality with Food#Style&Options or Food#Style&Options with Food#Prices. Besides, these experiments were carried out by the multi-task approach, therefore, this approach is more effective for sentences with two or three aspects.

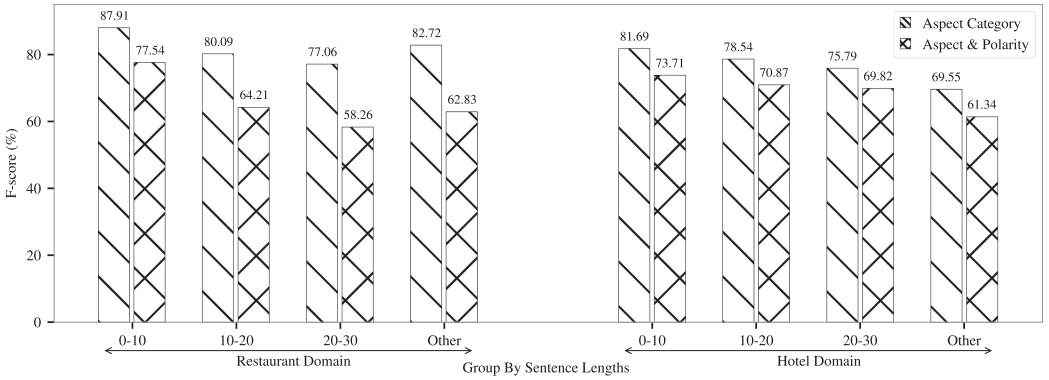

Fig. 4. The F1-score of two domains according to the length of sentence on the development set.

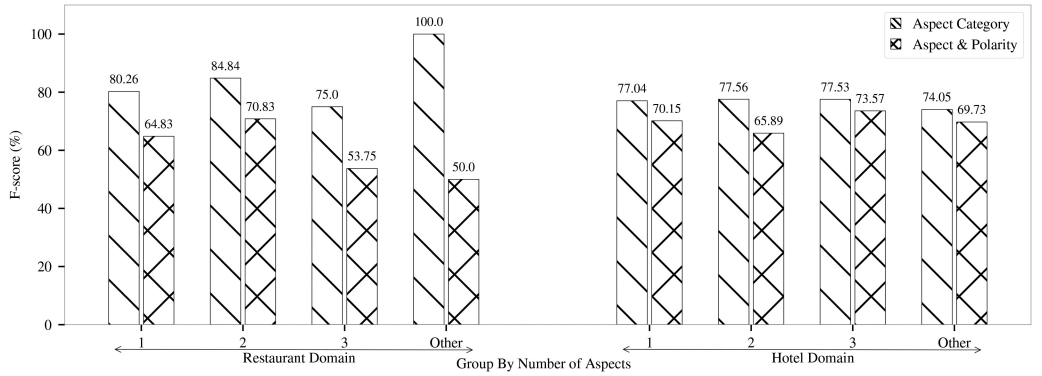

Fig. 5. The F1-score of two domains according to the number of aspects of each sentence on the development set.

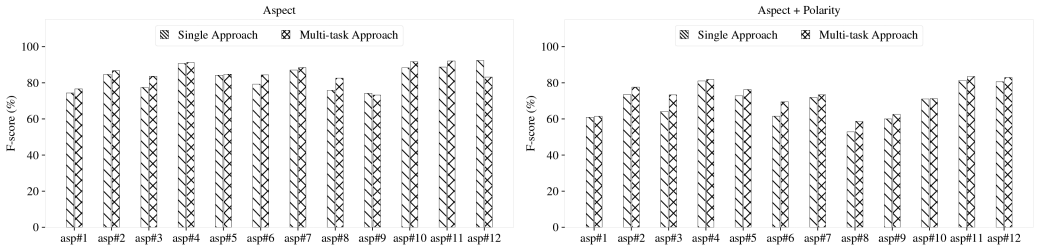

Fig. 6. Two bar charts show the comparison of BERT architecture based on two approaches to the development dataset for the restaurant domain. The left (a) and right (b) charts are the F1-scores of aspects for aspect category detection and aspect polarity task, respectively. The names of aspects are synchronized according to Table 1.

Figure 6 shows the different results of two approaches for aspects of the development dataset. As shown in Figure 6(a), we observe that the aspects of achieving an F1-score lower than 80% in both methods are Restaurant#General (asp#1), Restaurant#Miscellaneous (asp#3), Food#Prices (asp#5), Food#Style&Option (asp#6), Drink#Style&Option (asp#9). These aspects have a relatively high number of sentences on the training dataset, but the performance of two approaches based on various baseline architectures are still lower than others. Therefore, further researches should

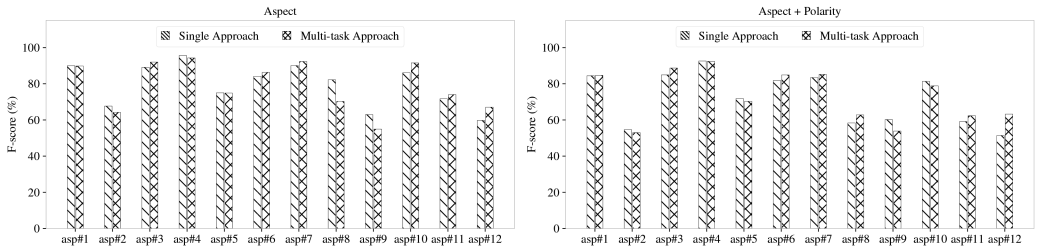

Fig. 7. The results of 12 first aspects for the hotel domain on the development dataset. The left (a) and right (b) charts are the F1-scores of first aspects for aspect category detection and aspect polarity task, respectively. The names of aspects are synchronized according to Table 6.

Table 11. The Experimental Results of Multi-task Approach Based on BERT Architecture on the Development and Challenge Sub-set(in %)

Dataset	Domain	Aspect Category			Aspect + Polarity		
		Precision	Recall	F1-score	Precision	Recall	F1-score
Development	Restaurant	86.20	83.67	84.92	73.68	71.51	72.58
	Hotel	80.60	75.76	78.10	74.54	70.06	72.23
Challenge Set	Restaurant	90.86	75.53	82.49	73.60	61.18	66.82
	Hotel	77.95	61.49	68.75	61.42	48.45	54.17

focus on improving the performance of these aspects to improve overall performance on our corpora.

As for the effectiveness of the two approaches, a further novel finding is that the multi-task approach gives better results in most of the aspects except for aspects such as Food#Quality(asp#5), Food#Style&Option (asp#6), Service#General (asp#12) for the aspect category detection task. It can be observed that the single-task approach is more effective for aspects with the amount of annotated data in the training dataset. For the main Aspect Polarity task, the single approach achieved the better F1-score for aspects (Restaurant#Prices(asp#2), Food#Quality(asp#5), Drink#Prices(asp#8), Service#General (asp#12)) than the multi-task approach. However, the difference between the two approaches is negligible except for the Drinks#Prices aspect. It is interesting to note that the predictive results of the single-task approach about sentiment of aspect is neutral: This demonstrates the bias of this aspect sentiment in the development dataset. Our analysis results demonstrate that the multi-task approach helps the model to know the relationships between different aspects in the data (e.g., the same entity or attribute) via training and sharing layers between aspects, which helps this approach based on different neural architecture to achieve better results than the single approach. We also show the results of 12 first aspects from Table 5 for the hotel domain in Figure 7. Similar to the results of the restaurant domain, the multi-task approach on the restaurant domain is still better than the single approach on most of the 12 aspects.

As mentioned above, we also extracted a challenging subset from the development set in which each sentence is assigned at least two different aspects with different polarities. In detail, we extracted two subsets of 109 and 59 sentences corresponding to the restaurant and hotel domains. Table 11 shows the experimental results on the development set and the challenging subsets for two domains. We can see that the performance of the best method on the challenging subset is lower than that on the development set related to the aspect with polarity task. The F1-score of

the restaurant and hotel domain decrease by −5.76% and −18.06% for the main task. This proves that the sentences with two different polarity labels greatly affect the overall results.

6 DISCUSSION

In this section, we will discuss the linguistic characteristics of Vietnamese compared to other resource-rich languages such as English and Chinese. After that, we conduct an experiment using machine translation technique to take advantage of the resource-rich language for predicting on Vietnamese reviews.

Compared with the English language, there are many differences between English and Vietnamese, such as language families [33] tones [25], tenses [23], sentence structures, and pronouns. Vietnamese includes a number of vocabulary originating from Chinese. About sentence structures, English sentence structures are usually opposed to Vietnamese sentence structures. For example, "A red apple" vs. "Một quả táo đỏ." A = một, red = đỏ, apple = quả táo. About tones, English is not a tonal language, but Vietnamese is. There are 6 tones in Vietnamese. When you say the term in various different tones, the context varies, as, for instance: Dứa (high up tone) = a pineapple, Dưa (flat tone) = a melon, Dừa (going down tone) = a coconut. About tenses, English has 12 tenses, however, there is really no tense in Vietnamese. English verbs can change their form with different pronouns or tenses. By way of example, she eats an apple/she ate an apple yesterday/she will eat an apple tomorrow. Vietnamese verbs do not change their form when you say various pronouns or the past, current, future. Alternatively, we can use đã/đang/sẽ add to verb to imply the past/present/future.

Like other languages in East Asia, both Chinese and Vietnamese are critical (isolating) languages [3]. Neither of them uses morphological marking of gender, case, tense, or number. The word order and function words in both languages convey grammatical relationships. The meaning will be changed accordingly, as word order or feature words are changed. Besides, its syntax conforms with the word order of the subject-verb-object and possesses noun classifier systems. Since every Chinese feature represents a significant unit, each syllable must be used separately as a meaningful unit of Vietnamese word building. Most Vietnamese words are bi-syllable, like Chinese. Chinese is written between words without blanks, while Vietnamese is written with two syllables, rather than words. Besides, the Vietnamese language is a prop-drop language, which implies that certain types of Vietnamese pronouns are removed when, in some sense, they are pragmatically infectious. Chinese also most frequently shows pro-drop characteristics. Differing in word order from Chinese, Vietnamese is head-initial, i.e., displaying modified-modifier ordering, but number and noun classifier being before the modified noun. Thus, for example, in Vietnamese grammar order, the Vietnamese language should not be Vietnamese language (Việt Nam tiếng) but language Vietnamese (tiếng Việt Nam)[49]. Therefore, in natural language processing and **machine translation (MT)** from Chinese or Vietnamese to different languages, in particular, the issue of word segmentation is often addressed primarily [38].

Can we solve the problem of low-resource language resources by machine translation? To answer this question, we conducted a small experiment on the aspect category detection as our validation task as follows: First, we test our experiment on the dataset, which uses Google Translate tool for creating training data from the English dataset. After translating from the SemEval 2016 dataset [26], we obtained 1,708 samples containing the aspect category. To the second model, we evaluate the performance of training on the Vietnamese annotated corpus with the same size. We train both models on a simple neural network and n-gram features. Then, we use two trained models to predict labels of the development dataset with 771 samples. The results show that the model trained on the Vietnamese dataset achieved 50.77% of F1-score, while the training model on the translated dataset only achieved 23.07% of F1-score. We manually inspected the case of

ineffective translation. The English sentence is "Moules was excellent, lobster ravioli was VERY salty!"; after translation, we got "Moules là tuyệt vời, tôm hùm ravioli mặn RẤT!" ("Moules are awesome, VERY salty ravioli lobsters!"). According to our above analysis on Vietnamese structure compared with English, especially when performing sentence translation, the semantic position is usually arranged in a "no" natural way. The correct translation should be "Moules rất ngon, tôm hùm ravioli RẤT mặn." ("Moules are awesome, lobster ravioli was VERY salty."). The data sources have increased significantly, but the model will still not effectively predict sentences used in real life, when the model can only learn nonsensical (or weird) sentences from translation, as in our example. This experiment demonstrated that the approach of using resource-rich language translation techniques through automated translation tools to make training dataset is not as effective as manual labelling on the language data itself. Another reason is that users' reviews often contain various abbreviations, spelling errors, and teen code; therefore, using translated corpus from other languages as training dataset is not effective. For the above reasons, we need to collect and annotate data for this problem on the users' comments.

7 CONCLUSION AND FUTURE WORK

In this article, we presented two benchmark corpora at sentence-level for aspect-based sentiment analysis in Vietnamese. Our corpora are annotated for two sub-tasks: Aspect Category Detection and Aspect Polarity Classification. Our corpora include 10,005 and 9,737 annotated sentences for the hotel domain and the restaurant domain, respectively. These are the biggest annotated corpora with high inner-annotation agreements for aspect-based sentiment analysis for the Vietnamese language. Our corpora will be published for the NLP community for free research purpose. In addition, we also explored the effectiveness of different single and multi-task machine learning models based on the neural architectures CNN, LSTM, BiLSTM, and BERT. The experimental results show important conclusions. The multi-task approach is better than the single approach for most of the architectures. The best architecture on our corpora is the BERT architecture for both tasks and both domains. We also analyzed the effectiveness of the best method according to sentence length and number of aspects mentioned. For future work, it would be interesting to apply the transfer learning strategy based on multi-task approach. Applying variant BERT-based architectures for our corpora is potential direction research. Another direction is to extend the corpora with annotations for the opinion target expression task.

ACKNOWLEDGMENTS

We would like to thank the reviewers for their helpful comments.

REFERENCES

[1] Plaban Kr. Bhowmick, Pabitra Mitra, and Anupam Basu. 2008. An agreement measure for determining inter-annotator reliability of human judgements on affective text. In *Proceedings of the Workshop on Human Judgements in Computational Linguistics*. Association for Computational Linguistics. 58–65.

[2] Xiao Chen, Changlong Sun, Jingjing Wang, Shoushan Li, Luo Si, Min Zhang, and Guodong Zhou. 2020. Aspect sentiment classification with document-level sentiment preference modeling. In *Proceedings of the 58th Meeting of the Association for Computational Linguistics*. 3667–3677.

[3] Yoon Mi Oh François Pellegrino Egidio and Marsico Christophe Coupé. 2013. A quantitative and typological approach to correlating linguistic complexity. *QITL-5* (2013), 71.

[4] Erfan Ghadery, Sajad Movahedi, Heshaam Faili, and Azadeh Shakery. 2019. MNCN: A multilingual Ngram-based convolutional network for aspect category detection in online reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 6441–6448.

[5] Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2018. Exploiting document knowledge for aspect-level sentiment classification. In *Proceedings of the 56th Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. 579–585. DOI: https://doi.org/10.18653/v1/P18-2092

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

[7] Mickel Hoang, Oskar Alija Bihorac, and Jacobo Rouces. 2019. Aspect-based sentiment analysis using bert. In *NEAL Proceedings of the 22nd Nordic Conference on Computional Linguistics (NoDaLiDa'19)*. Linköping University Electronic Press, Association for Computational Linguistics, Finland, 187–196.

[8] Robert Ireland and Ang Liu. 2018. Application of data analytics for product design: Sentiment analysis of online product reviews. *CIRP J. Manuf. Sci. Technol.* 23 (2018), 128–144.

[9] Jian Jin, Ying Liu, Ping Ji, and Hongguang Liu. 2016. Understanding big consumer opinion data for market-driven product design. *Int. J. Prod. Res.* 54, 10 (2016), 3019–3041.

[10] H. S. Le, T. V. Le, and T. V. Pham. 2015. Aspect analysis for opinion mining of Vietnamese text. In *Proceedings of the International Conference on Advanced Computing and Applications (ACOMP'15)*. IEEE, 118–123. DOI: https://doi.org/10.1109/ACOMP.2015.21

[11] Junjie Li, Haitong Yang, and Chengqing Zong. 2018. Document-level multi-aspect sentiment classification by jointly modeling users, aspects, and overall ratings. In *Proceedings of the 27th International Conference on Computational Linguistics*. 925–936.

[12] Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019. Exploiting BERT for end-to-end aspect-based sentiment analysis. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT'19)*. 34–41.

[13] Xin Li and Wai Lam. 2017. Deep multi-task learning for aspect term extraction with memory interaction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2886–2892. DOI: https://doi.org/10.18653/v1/D17-1310

[14] Bing Liu and Lei Zhang. 2012. *A Survey of Opinion Mining and Sentiment Analysis*. Springer US, Boston, MA, 415–463. DOI: https://doi.org/10.1007/978-1-4614-3223-4_13

[15] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2019. On the Variance of the Adaptive Learning Rate and Beyond. arxiv:cs.LG/1908.03265 (2019).

[16] Long Mai and Bac Le. 2018. Aspect-based sentiment analysis of Vietnamese texts with deep learning. In *Intelligent Information and Database Systems*, Ngoc Thanh Nguyen, Duong Hung Hoang, Tzung-Pei Hong, Hoang Pham, and Bogdan Trawiński (Eds.). Springer International Publishing, Cham, 149–158.

[17] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *CoRR* abs/1310.4546 (2013).

[18] Sajad Movahedi, Erfan Ghadery, Heshaam Faili, and Azadeh Shakery. 2019. Aspect category detection via topic-attention network. *CoRR*. http://arxiv.org/abs/1901.01183 (2019).

[19] Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. *Findings of EMNLP* (2020).

[20] Huyen Nguyen, Hung Nguyen, Quyen Ngo, Luong Vu, Vu Tran, Bach Ngo, and Cuong Le. 2019. VLSP shared task: Sentiment analysis. *J. Comput. Sci. Cyber.* 34, 4 (2019), 295–310. Retrieved from: http://vjs.ac.vn/index.php/jcc/article/view/13160

[21] M. Nguyen, T. M. Nguyen, D. Van Thin, and N. L. Nguyen. 2019. A corpus for aspect-based sentiment analysis in Vietnamese. In *Proceedings of the 11th International Conference on Knowledge and Systems Engineering (KSE'19)*. 1–5.

[22] T. P. Nguyen and A. C. Le. 2016. A hybrid approach to Vietnamese word segmentation. In *Proceedings of the IEEE RIVF International Conference on Computing Communication Technologies, Research, Innovation, and Vision for the Future (RIVF'16)*. IEEE, 114–119. DOI: https://doi.org/10.1109/RIVF.2016.7800279

[23] Nguyen Minh Nhut. 2020. An analysis of grammatical errors by Vietnamese learners of English. *Int. J. Adv. Res. Educ. Soc.* 2, 2 (2020), 23–34. Retrieved from: http://myjms.moe.gov.my/index.php/ijares/article/view/9652

[24] Thai-Hoang Pham and Phuong Le-Hong. 2017. End-to-end recurrent neural network models for Vietnamese named entity recognition: Word-level vs. character-level. *CoRR* abs/1705.04044 (2017).

[25] Ben Phạm and Sharynne McLeod. 2016. Consonants, vowels and tones across Vietnamese dialects. *Int. J. Speech-lang. Pathol.* 18, 2 (2016), 122–134. DOI: https://doi.org/10.3109/17549507.2015.1101162

[26] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*. Association for Computational Linguistics, 19–30. DOI: https://doi.org/10.18653/v1/S16-1002

[27] 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'15)*. Association for Computational Linguistics, 486–495. DOI: https://doi.org/10.18653/v1/S15-2082

[28] Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC Workshop on New Challenges for NLP Frameworks*. ELRA, 45–50. Retrieved from http://is.muni.cz/publication/884893/en

[29] Marzieh Saeidi, Guillaume Bouchard, Maria Liakata, and Sebastian Riedel. 2016. SentiHood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers (COLING'16)*. The COLING 2016 Organizing Committee, 1546–1556. Retrieved from https://www.aclweb.org/anthology/C16-1146

[30] Martin Schmitt, Simon Steinheber, Konrad Schreiber, and Benjamin Roth. 2018. Joint aspect and polarity classification for aspect-based sentiment analysis with end-to-end neural networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1109–1114. Retrieved from https://www.aclweb.org/anthology/D18-1139

[31] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases*, Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis (Eds.). Springer Berlin, 145–158.

[32] Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 380–385.

[33] Giang Tang. 2007. Cross-linguistic analysis of Vietnamese and English with implications for Vietnamese language acquisition and maintenance in the United States. *J. Southeast Asian Amer. Educ. Advanc.* 2, 1 (2007), 3.

[34] D. V. Thin, V. D. Nguye, K. V. Nguyen, and N. L. Nguyen. 2018. Deep learning for aspect detection on Vietnamese reviews. In *Proceedings of the 5th NAFOSTED Conference on Information and Computer Science (NICS'18)*. IEEE, 104–109.

[35] Dang Van Thin, Vu Nguyen, Nguyen Kiet, and Nguyen Ngan. 2019. A transformation method for aspect-based sentiment analysis. *J. Comput. Sci. Cyber.* 34, 4 (2019), 323–333. DOI: https://doi.org/10.15625/1813-9663/34/4/13162

[36] N. T. T. Thuy, N. X. Bach, and T. M. Phuong. 2018. Cross-language aspect extraction for opinion mining. In *Proceedings of the 10th International Conference on Knowledge and Systems Engineering (KSE'18)*. IEEE, 67–72.

[37] Khai Tran and Thi Phan. 2019. Deep learning application to ensemble learning—The simple, but effective, approach to sentiment classifying. *Appl. Sci.* 9, 13 (July 2019), 2760. DOI: https://doi.org/10.3390/app9132760

[38] Phuoc Tran, Dien Dinh, and Hien T. Nguyen. 2016. A character level based and word level based approach for Chinese-Vietnamese machine translation. *Computational intelligence and Neuroscience* 2016 (2016).

[39] Hai Wan, Yufei Yang, Jianfeng Du, Yanan Liu, Kunxun Qi, and Jeff Z. Pan. 2020. Target-aspect-sentiment joint detection for aspect-based sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 9122–9129.

[40] Jingjing Wang, Jie Li, Shoushan Li, Yangyang Kang, Min Zhang, Luo Si, and Guodong Zhou. 2018. Aspect sentiment classification with both word-level and clause-level attention networks. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. 4439–4445.

[41] Jingjing Wang, Changlong Sun, Shoushan Li, Jiancheng Wang, Luo Si, Min Zhang, Xiaozhong Liu, and Guodong Zhou. 2019. Human-like decision making: Document-level aspect sentiment classification via hierarchical reinforcement learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*. 5585–5594.

[42] Wenya Wang, Sinno Jialin Pan, and Daniel Dahlmeier. 2017. Multi-task coupled attentions for category-specific aspect and opinion terms co-extraction. *CoRR* abs/1702.01776 (2017).

[43] Michael Wojatzki, Eugen Ruppert, Sarah Holschneider, Torsten Zesch, and Chris Biemann. 2017. Germeval 2017: Shared task on aspect-based sentiment in social media customer feedback. *Proceedings of the GermEval* (2017), 1–12.

[44] Wei Xue and Tao Li. 2018. Aspect based sentiment analysis with gated convolutional networks. In *Proceedings of the 56th Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2514–2523.

[45] Wei Xue, Wubai Zhou, Tao Li, and Qing Wang. 2017. MTNA: A neural multi-task model for aspect category classification and aspect term extraction on restaurant reviews. In *Proceedings of the 8th International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, 151–156.

[46] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 1480–1489. DOI: https://doi.org/10.18653/v1/N16-1174

[47] Yichun Yin, Yangqiu Song, and Ming Zhang. 2017. Document-level multi-aspect sentiment classification as machine comprehension. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2044–2054. DOI: https://doi.org/10.18653/v1/D17-1217

[48] J. Yu, J. Jiang, and R. Xia. 2019. Global inference for aspect and opinion terms co-extraction based on multi-task neural networks. *IEEE/ACM Trans. Aud., Speech, Lang. Proc.* 27, 1 (2019), 168–177.

[49] Hai Zhao, Tianjiao Yin, and Jingyi Zhang. 2013. Vietnamese to Chinese machine translation via Chinese character as pivot. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC'13)*. 250–259.

[50] Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis C. M. Lau. 2015. A C-LSTM neural network for text classification. *CoRR* abs/1511.08630 (2015).

[51] Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2015. Representation learning for aspect category detection in online reviews. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*.

Received February 2020; revised December 2020; accepted January 2021