Span Detection for Aspect-Based Sentiment Analysis in Vietnamese

Kim Thi-Thanh Nguyen, Sieu Khai Huynh, Phuc Huynh Pham, Luong Luc Phan, Duc-Vu Nguyen, Kiet Van Nguyen
University of Information Technology, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
{18520963,18520348,18521260,18521073}@gm.uit.edu.vn
{vund, kietnv}@uit.edu.vn

Abstract

Aspect-based sentiment analysis is a challenging task that plays an essential role in natural language processing (NLP) and artificial intelligence. However, recent works only focused on aspect detection and sentiment classification but ignored the sub-task of detecting user opinion, which has immense potential in practical applications. In this paper, we present a new Vietnamese dataset (UIT-ViSD4SA) consisting of 35,396 human-annotated spans on 11,122 feedback comments for evaluating span detection for aspect-based sentiment analysis. Besides, we also propose a novel system using Bidirectional Long Short-Term Memory (BiLSTM) with a Conditional Random Field (CRF) layer (BiLSTM-CRF) for the span detection task in Vietnamese. The best result is a 62.76% $F1_{macro}$ for span detection using BiLSTM-CRF with embedding fusion of syllable embedding, character embedding, and contextual embedding from XLM-RoBERTa. In future work, span detection will be extended in many NLP tasks such as opinion mining, emotion recognition, complaint analysis, and constructive detection. Our dataset is freely available at `https://github.com/kimkim00/UIT-ViSD4SA` for research purposes.

1 Introduction

Typically, before buying an item or deciding to use a service, people tend to seek advice from their predecessors who purchased the item or used the service. With the explosive growth of e-commerce, more and more people find advice from websites, e-commerce sites, forums, or product review channels. Therefore, the number of reviews is increasing and becoming a valuable resource for customers and business. For customers, this data source provides information about products and helpful advice to help them avoid buying products or signing up for services that are not suitable for their personal needs. On the other hand, user reviews are also valuable information for businesses, and if used correctly and effectively, this data can help businesses improve product quality, accurately identify the target customers for each segment.

Aspect-based sentiment analysis (ABSA) (Hu and Liu, 2004) on user feedback is a challenging natural language processing task that attracts interest from both research and business (Jo and Oh, 2011; Kiritchenko et al., 2014; Chen et al., 2017). Given specific feedback about a product or service, the main task of ABSA is to detect what aspect is being discussed, then give sentiment analysis to the explored aspect. The ABSA problem can be divided into three basic tasks as follows: aspect detection, opinion target expression (OTE), sentiment polarity. In this paper, we focus on detecting the opinions of users based on aspects and their sentiment, which we call span detection for ABSA. Specifically, when a review is given *"Although staffs are nice, the phone is terrible!"*, the span detection for ABSA task aims to get two opinions *"staffs are nice"* and *"the phone is terrible"*, then classify these into right aspects also sentiment polarity.

The task is described as follows:

- **Input**: A customer feedback C for a smart-

phone that consists of n characters.

- **Output**: One or more spans of customer opinions are extracted directly from feedback C for each aspect. Each span is extracted from position i to position j such that $0 \leq i, j \leq n$ and $i \leq j$.

User interface contributes a significant part to the shopping experience on e-commerce sites. The user interface of these sites is becoming more convenient and user-friendly thanks to the help of ABSA techniques. If an e-commerce site adopts ABSA to their platform, customers can focus on corresponding reviews effectively by choosing the aspect-based sentiment text they care. On the other hand, the site owners can monitor their product and service qualities thanks to the help of ABSA. Popular Chinese e-commerce platforms such as Dianping, Taobao improve their user experience by deploying ABSA-based user interfaces. Therefore, the potential and importance of ABSA techniques for this area are immense. On the other hand, E-commerce sites present in Vietnam are still inferior in providing feedback to users. Most e-commerce platforms in Vietnam provide a simple feedback system: users leave their comments on the system along with a 5-star rating system like the one in Figure 1. Such systematic platforms include thegioididong[1], fptshop[2], shopee[3], tiki[4], and lazada[5]. Different from the rest, foody[6] (a restaurant review platform) allows users to leave a review, respond on a 10-point scale, and provides that score on several specific aspects (location, price, quality, service, and atmosphere). Therefore, we focus on the span detection for the ABSA problem, which not only detects aspects and their sentiment polarity but also detects specific opinions mentioned in comments; this will automatically provide a comprehensive and clear view of products and services.

To the best of our knowledge, current public Vietnamese datasets are constructed for only two popular tasks of ABSA (aspect extraction and sentiment

[1]https://www.thegioididong.com/
[2]https://fptshop.com.vn/
[3]https://shopee.vn/
[4]https://tiki.vn/
[5]https://www.lazada.vn/
[6]https://www.foody.vn/

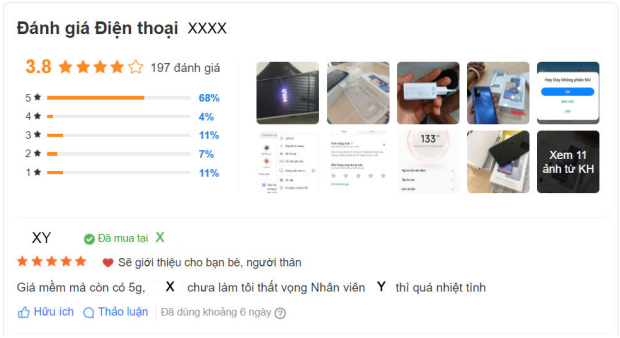

Figure 1: A feedback from an e-commerce site in Vietnam. The comment means "*(I) will recommended to friends and family: Good prices but still have 5g, X has never made me disappointed. Y staff is very enthusiastic.*" in English.

analysis), which limits further explorations of span detection. To addressing the problem and advancing the related research, this paper presents UIT-ViSD4SA, a benchmark Vietnamese smartphone feedback dataset for ABSA and span detection. All the feedback in UIT-ViSD4SA is collected from an e-commerce platform. There are 11,122 user comments, and each is manually annotated according to its spans towards ten fine-grained aspect categories with sentiment polarities. Figure 2 shows an annotated illustrative datapoint.

We have three main contributions summarized as follows:

- First and foremost, we create a benchmark Vietnamese feedback dataset toward span detection aspect category sentiment analysis, named UIT-ViSD4SA, including 35,396 spans on 11,122 real-world smartphone feedback comments annotated with ten aspect categories. The dataset is available freely for research purposes.

- Next, we propose an approach using BiLSTM-CRF with embedding fusion for span detection for ABSA in Vietnamese.

- Last but not least, we provide several case studies and future suggestions for span detection for ABSA in Vietnamese.

The rest of the paper is organized as follows. In Section 2, we present the related work. In Section

Input	Output
Máy đẹp, sangDESIGN#POSITIVE**, sd thì rất là ok**GENERAL#POSITIVE **máy mượt**PERFORMANCE#POSITIVE**. Pin sd cũng rất lâu mới hết, nhiều khi cả ngày và qua luôn ngày hôm sau mới sạc, sạc rất nhanh khoảng chừng 1 tiếng 5 phút là đầy rồi, ko lâu như iPhones mất gần 3 đến 4 tiếng đầy**BATTERY#POSITIVE. Chỉ sd để lướt web, facebook, youtube. Nghe nhạc rất hay đặc biệt là nghe bằng tai nghe AKG. **Rất xứng đáng với số tiền bỏ ra**GENERAL#POSITIVE., ***Beautifull phone, luxurious***DESIGN#POSITIVE***, use very ok***GENERAL#POSITIVE ***the machine is smooth***PERFORMANCE#POSITIVE***. The battery use also takes a long time to run out, sometimes it takes all day and the next day need to charge, very fast charging about 1 hour and 5 minutes is full, not as long as iphone takes nearly 3 to 4 hours to full***BATTERY#POSITIVE*. Only use to surf the web, facebook, youtube. Listening to music is very good, especially listening with AKG heedphones. ***Well worth the money spent***GENERAL#POSITIVE*.,*	0, 13, "DESIGN#POSITIVE" 15, 31, "GENERAL#POSITIVE" 32, 40, "PERFORMANCE#POSITIVE" 42, 175, "BATTERY#POSITIVE" 315, 346, "GENERAL#POSITIVE"

Figure 2: Examples illustrating spans for aspect-based sentiment analysis in Vietnamese.

3, we explain the data building process. The architecture of the approach is described in detail in Section 4. In Section 5, we implement a BiLSTM-CRF model to solve the problem and analysis to find the weakness of the method on our dataset. Finally, Section 6 draws conclusions and future work.

2 Related Work and Dataset

The SemEval shared-task published the ABSA datasets series that included user reviews from e-commerce sites, creating the foundation for much-related research (Li et al., 2019; Luo et al., 2020; Chen and Qian, 2020). The SemEval-2014 task 4 (SE-ABSA14) (Pontiki et al., 2014) dataset consists of restaurant and laptop reviews. The restaurant subset includes five aspects categories (i.e., Food, Service, Price, Ambience and Anecdotes/Miscellaneous) and four sentiment labels (i.e., Positive, Negative, Conflict and Neutral). The laptop subset was just annotated for aspect category detection and sentiment polarity classification. The SemEval-2015 Task 12 (SE-ABSA15) (Pontiki et al., 2015) dataset is built based on SE-ABSA14. SE-ABSA15 describes its aspect category as an entity type combined with an attribute type (e.g., Food#Style) and removes the Conflict sentiment label. The SemEval-2016 task-5 (SE-ABSA16) (Pontiki et al., 2016) dataset extended SE-ABSA15 to new domains such as Hotels, Consumer Electronics, Telecom, Museums, and other languages (Dutch, French, Russian, Spanish, Turkish, and Arabic).

Compared with rich resource languages such as English, Chinese, or Spanish, the number of high-quality Vietnamese datasets is still limited. Mai et al., (2018) collected smartphone reviews and annotated a Vietnamese ABSA corpus consisting of only 2,098 sentences for two tasks: OTE and sentiment classification. They presented a multi-task model using the sequence labeling scheme associated with bidirectional recurrent neural networks (BRNN) and conditional random field (CRF). The Vietnamese Language and Speech Processing (VLSP) community organized the first Vietnamese ABSA shared-task in 2018 (Nguyen et al., 2019a) provided a dataset composed of hotel and restaurant reviews collected from an online reviews platform. Unfortunately, the VLSP ABSA dataset inspired by SE-ABSA15 was only annotated for entity#atribute aspect category and its sentiment but ignored the OTE task. Nguyen et al.(2019b) presented a dataset on the same domains as the VLSP dataset, including only 7,828 reviews at document-level with seven aspect categories and five polarity sentiments for the same two tasks as the previous work: aspect extraction and sentiment classification. Dang et al., (2021) also built a dataset at the sentence level for the same domain as two previous works annotated with high inter-annotator agreements. To evaluate ABSA for mobile e-commerce, Phan et al., (2021) created a benchmark dataset (UIT-ViSFD) with 11,122 com-

Aspect	Definition
SCREEN	The users comment about the screen quality, size, colors, or display technology.
CAMERA	The comments mention the quality of a camera, vibration, delay, focus, or image colors.
FEATURES	The users refer to features, fingerprint sensor, wifi connection, touch or face detection of the phone.
BATTERY	The comments describes battery capacity or battery quality.
PERFORMANCE	The reviews describe ramming capacity, processor chip, performance using, or smoothness of the phone.
STORAGE	The comments mention storage capacity, the ability to expand capacity through memory cards.
DESIGN	The reviews refer to the style, design, or shell.
PRICE	The comments present the price of the phone.
GENERAL	The reviews of customers generally comment about the phone.
SER&ACC[7]	The comments mention sales service, warranty, or review of accessories of the phone.

[7] SER&ACC is short for SERVICE and ACCESSORIES.

Table 1: The full list of ten aspects and their brief definitions (Phan et al., 2021).

ments based on a strict annotation scheme. Furthermore, they developed a social listening system that can help motivate for application of ABSA in Vietnamese e-commerce sites.

About span-related methods for ABSA task, Hu et al., (2019) proposed a span-based extract-then-classify framework. Their model directly extracted multiple opinion targets from the review under the supervision of target span boundaries, and then corresponding sentiment polarities are classified using their span representations. This work was inspired by advances in machine reading comprehension and question answering (Seo et al., 2018; Xu et al., 2018), whose aim is to extract a continuous span of text from the document as an answer for a question. (Rajpurkar et al., 2016; Nguyen et al., 2020). Xu et al., (2020) presented a neat and effective multiple CRFs based structured attention model capable of extracting aspect-specific opinion spans. The sentiment polarity of the target is then classified based on the extracted opinion features and contextual information.

3 Dataset Creation and Analysis

Based on the benchmark dataset proposed by Phan et al., (2021), we develop a new dataset for span detection for ABSA in Vietnamese. The three-phase creation process of our dataset is explained as follows. To begin with, we edit and supplement the annotation guidelines from (Phan et al., 2021) for annotators to understand the span definition and how to annotate correctly (see Section 3.1). Annotator team members are trained with the guidelines at the same time as annotating the guidelines-training data to ensure that the F1-score in the training process reaches over 80%. Finally, the team members annotate data independently to maximize annotation performance (see Section 3.2). Besides, we provide an analysis of the dataset that helps experts have an overview of this dataset (see Section 3.3).

We utilize the UIT-ViSFD dataset(Phan et al., 2021) collected from an e-commerce website for smartphones in Vietnam, which allows customers to write fine-grained reviews of a smartphone they have purchased or used. In the reviews, users give positive, neutral, or negative opinions on many aspects explicitly or implicitly, such as camera, price, battery, service, etcetera. The dataset includes 11,122 feedback with four attributes: *comment, n_star, date_time, and label*. Table 1 summarizes ten aspects in the guidelines, and each aspect has one of three sentiment polarities (positive, negative, and neutral).

3.1 Span Definition and Annotation Guidelines

Following the annotation guidelines proposed by Phan et al., (2021), we add some definitions and rules to form the core of data construction. We reuse the ten predefined aspect categories as in Table 1, with each aspect category mentioned in the review, the sentiment polarity for the aspect is labeled as POSITIVE, NEUTRAL, or NEGATIVE. The span is defined as the shortest span containing the opinions of the user about the aspect category. With ten predefined aspects, annotators are asked to annotate spans towards aspect categories with sentiment polarities of each review. Suppose a review is given, when a span is discovered within the review either explicitly or implicitly, the aspect category with sentiment polarity of that span is labeled

as aspect#polarity as in Figure 2.

3.2 Annotation Process

Three phases of annotation are conducted as follows. To begin with, we train annotators with the guidelines and randomly take about 30-70 reviews in the dataset to annotate, then calculate F1-core per review for those annotated data. For disagreement cases, annotators decides the final label by discussing and having a voting poll, then clarify the vague term or supplement the unknown term in the guidelines. Annotators team members trained four rounds to obtain a high F1-score above 80% before performing data annotation independently. Figure 3 shows the F1-score during training phases.

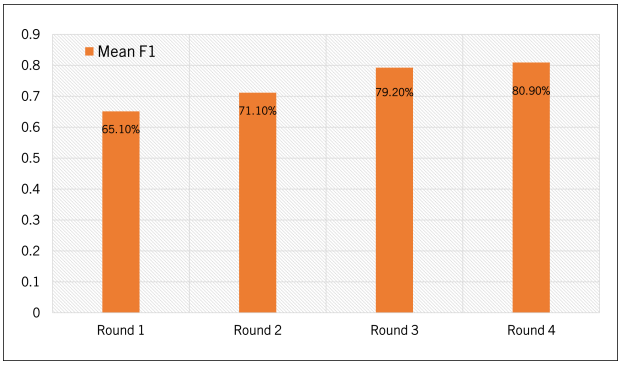

Figure 3: Results for four rounds of measurement of F1-score.

An annotation is a triple (d, l, o), where d is a document id, l a label, and o is a list of start-end character offset tuples. An annotator i contributes a (multi)set A_i of (token) annotations. We compute (1) for each 2-combination of annotators and report arithmetic mean of F1 across all these combinations (Hripcsak and Rothschild, 2005). Grouping annotations by documents or labels allows us to calculate F1 per document or label.

$$F1_{ij} = \frac{2 \times |A_i \cap A_j|}{A_i + A_j} \quad (1)$$

Finally, our dataset is divided randomly into three sets: training (Train), development (Dev), and test (Test) in the ratio 7:1:2.

3.3 Dataset Analysis

Figure 4 presents the distribution of ten aspect categories in our dataset UIT-ViSD4SA. People tend to give a smartphone an overall rating, with 22.76% of reviews mentioning GENERAL aspect. Users frequently pay great attention to aspects related to porpular needs, such as PERFORMANCE, BATTERY, FEATURES, and CAMERA.

The statistics of our dataset are presented in Table 2. Our dataset includes 35,396 spans over 11,122 comments. Through the fundamental analysis, the dataset has an imbalanced distribution of the sentiment labels. The positive labels account for the most significant number, followed by the negative and neutral labels last. On average, the reviews have three spans, with each span being about 32 characters long. We hope our dataset will open a new shared-task about ABSA and help apply the ABSA span detection technique for e-commerce systems.

4 Our Approach

For the baseline evaluation, we consider the span detection task as a sequence labeling problem at the syllable level. We employ a BiLSTM-CRF model (Huang et al., 2015) with embedding fusion to solve the task. The BiLSTM-CRF model comprises three layers: token embedding layer giving contextualized vector representation of input sequence, passed into the BiLSTM-CRF sequence labeler as depicted in Figure 5.

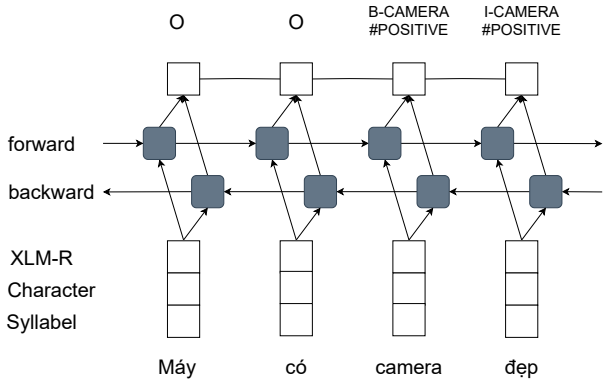

Figure 5: BiLSTM-CRF network with embedding layers (the example feedback means *"This phone has a good camera"* in English).

4.1 Embedding Fusion Layer

The embedding layer takes as input a sequence of N tokens $(x_1, x_2, .., x_N)$, and output a fixed-dimensional vector representation of each token

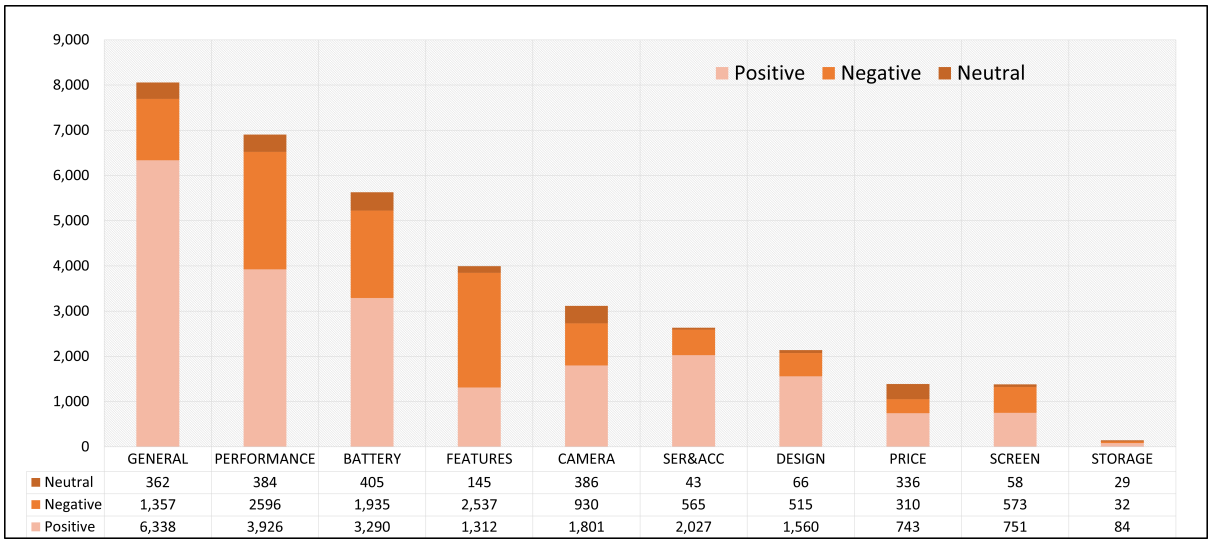

Figure 4: The distribution of 10 fine-grained aspect categories.

Set	Comment	Average aspect per comment	Average span length	Positive	Negative	Neutral	Total span
Train	7,784	3.2	32.6	15,356	7,793	1,560	
Dev	1,113	3.1	32.4	2,110	1,144	241	35,396
Test	2,225	3.2	32.5	4,266	2,269	413	

Table 2: The overview statistics of UIT-ViSD4SA dataset.

$(e_1, e_2, ..., e_N)$. We use an embedding fusion of syllable embedding (Nguyen et al., 2017), character embedding (CharLSTM), contextual embedding from XLM-RoBERTa (Conneau et al., 2020).

4.2 Bidirectional Long Short-Term Memory (BiLSTM)

A long-short term memory network (LSTM) is a special type of Recurrent neural network (RNN) introduced by Hochreiter et al., (1997), which can capture a long-distance semantic relationship by maintaining a memory cell store context information. LSTMs do not suffer from vanishing and exploding gradient problems. The LSTM is equipped with a memory cell with an adaptive adjustment mechanism that adjusts information to be added to or removed from the cell. The memory cell is continuously updated during encryption, and the information rate is determined by three kernel gates, including input, forget and output. In terms of formality, the encryption process at the time step *t* is performed as follows:

$$i_t = \sigma(W_{hi}h_{t-1} + W_{ei}e_t^w + b_i) \quad (2)$$

$$f_t = \sigma(W_{hf}h_{t-1} + W_{ef}e_t^w + b_f) \quad (3)$$

$$\widetilde{c_t} = tanh(W_{hc}h_{t-1} + W_{ec}e_t^w + b_c) \quad (4)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \widetilde{c_t} \quad (5)$$

$$o_t = \sigma(W_{ho}h_{t-1} + W_{eo}e_t^w + b_o \quad (6)$$

$$h_t = o_t \odot tanh(c_t) \quad (7)$$

where c_t, i_t, f_t, and o_t represent the memory cell, input gate, forget gate and output gate respectively. e_t^w and h_t donate the word embedding vector and hidden state vector at time t. Both σ and $tanh$ are the activation functions, and $\odot$ represents the element-wise product. $W*$ and $b*$ are network parameters that donate the weight matrices and bias vectors. Although LSTM can solve the long-distance dependency problem, it still loses some semantic information due to the sequential encoding way of LSTM. For example, h_t only contains the semantic information before time step t. Therefore, a BiLSTM is needed to model both the forward and

backward context information as in equation (8,9), and the two hidden states are concatenated to obtain the final output as equation (10):

$$\overrightarrow{h_t} = F(e_t^w, \overrightarrow{h_{t-1}}) \quad (8)$$

$$\overleftarrow{h_t} = F(e_t^w, \overleftarrow{h_{t-1}}) \quad (9)$$

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \quad (10)$$

4.3 Conditional Random Fields (CRF)

Conditional Random Fields (CRF) (Lafferty et al., 2001) is a sequence modeling framework that brings in all the advantages of MEMMs models (McCallum et al., 2000; Ratnaparkhi, 1996) while also solving the label bias problem. CRF directly connected the inputs and outputs, unlike LSTM and BiLSTM networks where memory cells/recurrent components are employed. Given a training dataset $D = (x^1, y^1), ..., (x^N, y^N)$ of N data sequences to be labeled x^i and their corresponding label sequences y^i, CRF maximizes the conditional log-likelihood of label sequences based on the data sequences as shown as follow:

$$L = \sum_{i=1}^{N} log(P(y^i|x^i)) - \sum_{k=1}^{K} \frac{\lambda_k^2}{2\sigma^2} \quad (11)$$

5 Experiments and Results

5.1 Experimental Settings

Following the IOB format (short for inside, outside, beginning), our dataset is converted with data containing only aspect labels (SCREEN, BATTERY, CAMERA, etcetera.), sentiment labels only (POSITIVE, NEUTRAL, and NEGATIVE), and data containing both aspect and sentiment labels (SCREEN#POSITIVE, BATTERY#NEGATIVE, etcetera.) to evaluate our approach comprehensively. Our word embeddings have three parts: syllable (1), character (2), and contextual from XLM-R(3), with an embedding dimension of 100. We set the hidden layers of LSTM as 400, the dropout rate as 0.33, and the batch size as 5,000 with 30 epochs for training. All experiments are conducted on a single NVIDIA T4 GPU card.

5.2 Evaluation Metrics

In this paper, we use three evaluation metrics: Precision, Recall, and F1-score. A predicted span is correct only if it exactly matches the gold standard span. We calculate these evaluation metrics on both the micro and macro averages to gain a wide view.

5.3 Experimental Results

Table 3 presents performances of the BiLSTM-CRF model with three types of embedding fusion on the aspect, polarity, aspect#polarity span detection. According to our results, we can see that concatenate three embedding layers (syllable, character, and bert-based embedding) have significantly better performance than just one or two embedding layers. In particular, syllabel+char+XLMRlarge achieves the best $F1_{macro}$ of 62.76%, 49.77%, and 45.70% for aspect, polarity, and aspect#polarity, respectively, whereas the model with just syllable embedding layer shows the lowest performances. On the other hand, our method tends to be less efficient with labels which consist of polarity, in which polarity task reach 49.77% $F1_{macro}$ while aspect#polarity task gets 45.70% $F1_{macro}$.

Detailed results per class of each task are shown in Tables 4, 5, and 6 (with aspect#polarity label, we only show F1-score). For aspect task, only two aspects have a high F1-score above 70% (CAMERA and BATTERY) while the rest range from 60-70%, especially F1-score of PRICE and STORAGE is relatively low (below 50%). With the polarity task, the result is descending with the order POSITIVE, NEGATIVE, NEUTRAL. The result of aspect#polarity can be considered the sum of the two previous tasks: previous high-performing aspects labels combined with positive give the highest result. This result explains the imbalance in the sentiment labels as shown in Figure 4 and Table 2 (NEUTRAL labels only cover 6.25% of our dataset). Our approach generally gets better performance when detecting span for aspect than polarity and aspect#polarity. However, the ability of all models to detect span for all types of label form is still limited (F1-score below 80%), which will be exploited in future work.

5.4 Case Study

Figure 6 shows several cases predicted by the BiLSTM-CRF model. After reviewing the cases, we found that the model commits three common types of errors that can not detect spans, misclassify the

System	$\mathbf{P}_{Micro}$	$\mathbf{R}_{Micro}$	$\mathbf{F1}_{Micro}$	$\mathbf{P}_{Macro}$	$\mathbf{R}_{Macro}$	$\mathbf{F1}_{Macro}$
Aspect (syllable)	64.55	60.86	62.65	62.76	57.28	59.74
Aspect (syllable + char)	63.78	62.11	62.93	61.64	58.91	60.21
Aspect (syllable + char + XLM-R_{Base})	65.63	65.15	65.39	62.88	61.62	62.17
Aspect (syllable + char + XLM-R_{Large})	64.96	66.85	**65.89**	62.00	63.56	**62.76**
Polarity (syllable)	52.36	50.10	51.20	46.71	38.37	41.05
Polarity (syllable + char)	52.12	51.00	51.55	44.44	38.79	40.68
Polarity (syllable + char + XLM-R_{Base})	54.88	55.91	55.39	46.87	46.39	46.57
Polarity (syllable + char + XLM-R_{Large})	56.89	59.78	**58.30**	49.00	50.60	**49.77**
Aspect-polarity (syllable)	61.87	54.55	57.98	48.77	34.27	37.64
Aspect-polarity (syllable + char)	59.51	57.56	58.52	43.66	37.53	39.30
Aspect-polarity (syllable + char + XLM-R_{Base})	60.71	61.62	61.16	46.18	43.42	44.37
Aspect-polarity (syllable + char + XLM-R_{Large})	61.78	62.99	**62.38**	46.84	45.46	**45.70**

Table 3: The overall experimental results.

Aspect	Precision	Recall	F1-score
BATTERY	71.04	73.58	72.29
CAMERA	75.09	77.82	**76.43**
DESIGN	68.13	70.66	69.37
FEATURES	58.76	59.34	59.05
GENERAL	64.74	68.90	66.76
PERFORMANCE	62.37	63.11	62.74
PRICE	46.72	47.98	47.35
SCREEN	65.83	68.70	67.23
SER&ACC	65.18	61.83	63.46
STORAGE	45.16	46.67	45.90

Table 4: Result per class for only aspect label.

Sentiment	Precision	Recall	F1-score
NEGATIVE	47.05	47.56	47.30
NEUTRAL	36.57	35.97	36.26
POSITIVE	63.52	68.50	**65.92**

Table 5: Result per class for only sentiment polarity label.

Aspect	Negative	Neutral	Positive
BATTERY	54.62	44.07	**78.40**
CAMERA	58.97	55.65	77.54
DESIGN	46.15	00.00	75.75
FEATURES	50.73	22.22	68.11
GENERAL	52.12	52.73	67.87
PERFORMANCE	45.87	24.19	70.84
PRICE	32.69	15.05	52.63
SCREEN	48.62	46.15	71.13
SER&ACC	22.56	00.00	72.17
STORAGE	15.38	00.00	57.14

Table 6: F1-score per class for aspect#polarity label.

sentiment polarity, and detect the wrong boundary of spans. As observed in the first sentence, both three types of models can not detect the span *"there's some sound from the speaker"*. With the cases of misclassification, we found that many cases of this mistake contained English loanwords. For example, in comment 2, the span *"Really like the dark mode"* is about the interface, and we annotate it as PERFORMACNE#POSITIVE. However, the model can understand it and classify it as CAMERA (aspect label model) or FEAUTURE#POSITIVE (aspect#polarity label model). This feature needs attention and research in future studies because the Vietnamese language feature (especially in technology) often includes many loanwords with meanings that can be similar or different from the original language. Besides, the polarity model incorrectly predicts the target span by detecting the whole span *"the screen is clear, play game phone is warm, but noise speaker"* as a NEUTRAL span. This mistake can be blamed on the way we train the model just with polarity label, which makes it difficult for the model. The proof for this argument is that the model with the label aspect#polarity can detect the boundary of spans better than the polarity model.

6 Conclusion and Future Work

This paper presented UIT-ViSD4SA, which is a new dataset for span detection on aspect-based sentiment analysis and consists of over 35,000 human-annotated spans on 11,122 comments for mobile e-commerce. Each feedback is manually annotated ac-

	Gold labels	Aspect prediction	Polarity prediction	Aspect#polarity prediction
1	tôi cảm thấy, **loa có tiếng gì đó phát ra**$_{\text{FEATURES\#NEGATIVE}}$, mặc dù k chạm vào điện thoại.còn lại **in trâu**$_{\text{BATTERY\#POSITIVE}}$, **màn nét**$_{\text{SCREEN\#POSITIVE}}$, **chơi game ấm**$_{\text{PERFORMANCE\#NEGATIVE}}$, nhưng **loa dè**$_{\text{FEATURES\#NEGATIVE}}$ *i feel,* ***there're some sound from the speaker****, even though I don't touch the phone.the rest is* ***battery last long****,* ***the screen is clear****,* ***play game phone is warm****, but* ***noisy speaker***	màn nét$_{\text{SCREEN}}$ chơi game ấm$_{\text{PERFORMANCE}}$ loa dè$_{\text{FEATURES}}$	màn nét, chơi game ấm, nhưng loa dè$_{\text{NEUTRAL}}$X	in trâu$_{\text{BATTERY\#POSITIVE}}$ màn nét$_{\text{SCREEN\#POSITIVE}}$ chơi game ấm$_{\text{PERFORMANCE\#POSITIVE}}$X loa dè$_{\text{FEATURES\#NEGATIVE}}$
2	Sử dụng hơn 3 tháng thấy **máy rất tốt**$_{\text{GENERAL\#POSITIVE}}$, **dùng 2 ngày mới sạc lần , lần sạc 2-3 tiếng là đầy**$_{\text{BATTERY\#POSITIVE}}$. **Rất thích dark mode**$_{\text{PERFORMANCE\#POSITIVE}}$. *Using more than 3 months find that* ***the device is really goo****,* ***using till 2 days to need to charge , take 2-3 hours to full****.* ***Really like the dark mode****.*	máy rất tốt$_{\text{GENERAL}}$ dùng 2 ngày mới sạc lần , lần sạc 2-3 tiếng là đầy$_{\text{BATTERY}}$ Rất thích dark mode $_{\text{CAMERA}}$X	máy rất tốt$_{\text{POSITIVEL}}$ dùng 2 ngày mới sạc lần , lần sạc 2-3 tiếng là đầy$_{\text{POSITIVE}}$ Rất thích dark mode $_{\text{POSITIVE}}$	máy rất tốt$_{\text{GENERAL\#POSITIVE}}$ dùng 2 ngày mới sạc lần , lần sạc 2-3 tiếng là đầy$_{\text{BATTERY\#POSITIVE}}$ Rất thích dark mode $_{\text{FEATURES\#POSITIVE}}$X

Figure 6: Case study. The spans are bold with aspects and their polarities are given as subscripts. Incorrect predictions are marked with X.

cording to its spans towards ten fine-grained aspect categories with their sentiment polarities. BiLSTM-CRF uses an embedding fusion of syllable, character, and contextual embedding, which had the highest 62.76% $F1_{macro}$ for span detection on aspect, 49.77% $F1_{macro}$ for span detection on polarity, and 45.70% $F1_{macro}$ for span detection on aspect#polarity, respectively. In general, the performances are relatively not high and challenging for further machine learning-based models. We hope the release of UIT-ViSD4SA could motivate the development of machine learning models and applications.

In future work, we give several directions: (1) Inspired by Yuan et al., (2020), multilingual pre-trained language models can be used for enhancing span boundary detection. (2) Improving the performance of this task can be used with approaches based on machine comprehension reading and other approaches (Hu et al., 2019; Xu et al., 2020). (3) Inspired by Xu et al., (2019), we can develop a review reading comprehension for Vietnamese based on our dataset. (4) Span detection is a challenging task that can motivate various works of many topic such as: constructive analysis (Fujita et al., 2019; Nguyen et al., 2021a), emotion analysis (Sosea and Caragea, 2020; Ho et al., 2019), complaint analysis (Preoţiuc-Pietro et al., 2019; Nguyen et al., 2021b), and opinion mining (Nguyen et al., 2018; Jiang et al., 2019).

References

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

Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 452–461, Copenhagen, Denmark, September. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Soichiro Fujita, Hayato Kobayashi, and Manabu Okumura. 2019. Dataset creation for ranking constructive news comments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2619–2626.

Vong Anh Ho, Duong Huynh-Cong Nguyen,

Danh Hoang Nguyen, Linh Thi-Van Pham, Duc-Vu Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2019. Emotion Recognition for Vietnamese Social Media Text. In *International Conference of the Pacific Association for Computational Linguistics*, pages 319–333. Springer.

S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

G. Hripcsak and A. Rothschild. 2005. Technical brief: Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association : JAMIA*, 12 3:296–8.

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

Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheng Li, and Yiwei Lv. 2019. Open-Domain Targeted Sentiment Analysis via Span-Based Extraction and Classification. In *Proceedings of ACL*, pages 537–546. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6280–6285.

Yohan Jo and Alice H. Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, page 815–824. Association for Computing Machinery.

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. 2014. NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 437–442, Dublin, Ireland. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289. Morgan Kaufmann Publishers Inc.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019. A unified model for opinion target extraction and target sentiment prediction. In *AAAI*.

Huaishao Luo, Lei Ji, Tianrui Li, Daxin Jiang, and Nan Duan. 2020. GRACE: Gradient harmonized and cascaded labeling for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 54–64. Association for Computational Linguistics.

Long Mai and Bac Le. 2018. Aspect-based sentiment analysis of vietnamese texts with deep learning. In *Asian Conference on Intelligent Information and Database Systems*, pages 149–158. Springer.

A. McCallum, Dayne Freitag, and Fernando C Pereira. 2000. Maximum entropy markov models for information extraction and segmentation. In *ICML*.

Dat Quoc Nguyen, Thanh Vu, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2017. From word segmentation to POS tagging for Vietnamese. In *Proceedings of the Australasian Language Technology Association Workshop 2017*, pages 108–113.

Huyen TM Nguyen, Hung V Nguyen, Quyen T Ngo, Luong X Vu, Vu Mai Tran, Bach X Ngo, and Cuong A Le. 2018. Vlsp shared task: sentiment analysis. *Journal of Computer Science and Cybernetics*, 34(4):295–310.

Huyen Nguyen, Hung Nguyen, Quyen Ngo, Luong Vu, Vu Tran, Ngo Xuan Bach, and Cuong Le. 2019a. Vlsp shared task: Sentiment analysis. *Journal of Computer Science and Cybernetics*, 34:295–310, 01.

Minh-Hao Nguyen, Tri Minh Nguyen, Dang Van Thin, and Ngan Luu-Thuy Nguyen. 2019b. A corpus for aspect-based sentiment analysis in vietnamese. In *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*, pages 1–5.

Kiet Nguyen, Vu Nguyen, Anh Nguyen, and Ngan Nguyen. 2020. A vietnamese dataset for evaluating machine reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2595–2605.

Luan Thanh Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021a. Constructive and toxic speech detection for open-domain social media comments in vietnamese. In *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices*, pages 572–583. Springer International Publishing.

Nhung Thi-Hong Nguyen, Phuong Ha-Dieu Phan, Luan Thanh Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021b. Vietnamese complaint detection on e-commerce websites. *arXiv preprint arXiv:2104.11969*.

Luong Luc Phan, Phuc Huynh Pham, Kim Thi-Thanh Nguyen, Tham Nguyen, Sieu Khai Huynh, Luan Thanh Nguyen, Tin Van Huynh, and Kiet Van Nguyen. 2021. Sa2sl: From aspect-based sentiment analysis to social listening system for business intelligence. In *KSEM*.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In *Proceedings of ACL*, pages 27–35. Association for Computational Linguistics.

Maria Pontiki, Dimitrios Galanis, Harris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 Task 12: Aspect based sentiment analysis. In *Proceedings of SemEval*, pages 486–495.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 Task 5: Aspect based sentiment analysis. In *Proceedings of SemEval*, pages 19–30.

Daniel Preoţiuc-Pietro, Mihaela Gaman, and Nikolaos Aletras. 2019. Automatically identifying complaints in social media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5019. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Conference on Empirical Methods in Natural Language Processing*.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2018. Bidirectional attention flow for machine comprehension.

Tiberiu Sosea and Cornelia Caragea. 2020. Canceremo: A dataset for fine-grained emotion detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8892–8904.

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).

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

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

Lu Xu, Lidong Bing, Wei Lu, and Fei Huang. 2020. Aspect Sentiment Classification with Aspect-Specific Opinion Spans. In *Proceedings of EMNLP*, pages 3561–3567. Association for Computational Linguistics.

Fei Yuan, Linjun Shou, Xuanyu Bai, Ming Gong, Yaobo Liang, Nan Duan, Yan Fu, and Daxin Jiang. 2020. Enhancing answer boundary detection for multilingual machine reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 925–934.