VNU Journal of Science:
Computer Science and Communication Engineering

Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Original Article

Aspect-Category based Sentiment Analysis with Unified Sequence-To-Sequence Transfer Transformers

Dang Van Thin, Ngan Luu Thuy Nguyen[*]

University of Information Technology
Quarter 6, Linh Trung Ward, Thu Duc District, Ho Chi Minh City, Vietnam
Vietnam National University Ho Chi Minh City, Ho Chi Minh City, Vietnam

Received 06 February 2023
Revised 08 May 2023; Accepted 01 June 2023

Abstract: In recent years, Aspect-based sentiment analysis (ABSA) has received increasing attention from the scientific community for Vietnamese language. However, most previous studies solved various subtasks in ABSA based on machine learning, deep learning and transformer-based architectures in a classification way. Recently, the release of pre-trained sequence-to-sequence brings a new approach to address the ABSA as a text generation problem for Vietnamese ABSA tasks. In this paper, we formulate the Aspect-category based sentiment analysis task as the conditional text generation task and investigate different unified generative transformer-based models. To represent the labels in a natural sentence, we apply a simple statistical method and observation of the commenting style. We conduct experiments on two benchmark datasets. As a result, our model achieved new state-of-the-art results with the micro F1-score of 75.53% and 86.60% for the two datasets with different levels for the restaurant domain. In addition, our experimental results achieved the best score for the smartphone domain with the macro F1-score of 81.10%.

Keywords: Aspect-based sentiment analysis, unified model, Vietnamese language, pre-trained language models.

1. Introduction

Aspect-based sentiment analysis (ABSA) is a fine-grained sentiment analysis toward specific aspects. This problem consists of different sub-tasks, two of which are Aspect Category Detection (ACD) and Sentiment Analysis for the

[*] Corresponding author.
E-mail address: ngannlt@uit.edu.vn

https://doi.org/10.25073/2588-1086/vnucsce.662

Aspect Category. The purpose of the Aspect Category Detection task is aim to detect the list of aspect categories mentioned in the review. In contrast, the Sentiment Analysis for the Aspect Category task is to predict the sentiment polarity class towards each category. For instance, given a simple example for the restaurant domain, "*The food is delicious but its price was expensive*", the expected outputs for the aspect category detection are "*Food#Quality*" and "*Food#Prices*". The corresponding sentiments are "positive" for "*Food#Quality*" category and "negative" for "*Food#Prices*", respectively. In this paper, we focus on the compound task of the two sub-tasks, the input is a review, and the output is the list of aspect categories and corresponding sentiments (denoted as Aspect-Category based Sentiment Analysis - ACSA).

Many research efforts have solved the ACSA task as the classification task based on the power of machine learning algorithms. One of the straightforward methods is to identify the mentioned aspect categories and then predict the sentiment class corresponding to each aspect category [1-3]. However, this approach highly depends on the performance of detecting aspect category components. Moreover, the relation between the two tasks is not utilized in this approach. For that reason, early studies demonstrated that the multi-task learning framework had shown effectiveness for this task [4-7]. Recently, with the development of generative text models such as T5 [8], BART [9], Liu et al. [10] addressed the ACSA tasks based on the sequence-to-sequence modelling. The experimental results showed that this approach outperforms previous studies.

Inspired by recent success in solving the Natural Language Processing tasks as text generation paradigm [10-12], we present a unified sequence-to-sequence transformer-based in an end-to-end manner to tackle the ACSA task in Vietnamese benchmark datasets. The main idea of the text generation approach is to represent the output as a natural language sentence and take advantage of pre-trained generative models. However, most Vietnamese datasets use English labels to represent the output instead of Vietnamese. To address this challenge, we applied a simple statistical method and observed the commenting style of users to convert the output labels to natural sentences.

2. Related work

For the Vietnamese ABSA research, there have been a lot of studies in recent years. First, the work of Huyen et al. [1] organized a shared-task VLSP challenge for the ABSA task in 2018. In the competition, two document-level datasets (denoted as VLSP datasets) provided for the participants. There were many methods have been tested, however, most of the competitor's approaches are based on traditional machine learning algorithms combined with handcraft features. At the competition, Thin et al. [2] presented a transformation approach by converting the ACSA task as the multiple binary classification problem that achieved the best performance on two benchmarks. Inspired by their work, Thin et al. [13] explored the performance of Deep Convolutional Neural Network (DCNN) to analyze Vietnamese comments on the ACD task of two VLSP datasets.

Based on the guidelines of VLSP datasets, the authors [6] provided two large sentence-level datasets for the restaurant and hotel reviews in Vietnamese. The datasets were released to promote the growth of the NLP community for the Vietnamese ABSA research. The authors also conducted massive experiments on the multi-task approach compared with the single approach based on machine learning, deep learning, and transformer-based models. The experimental results concluded that the PhoBERT model combined with the multi-task approach outperformed the machine learning and deep learning architectures and the single approach. Phan et al. [14] provided a UIT-

ViSFD (Vietnamese Smartphone Feedback Dataset) dataset consisting of 11,122 labelled mobile e-commerce comments. They also used this dataset to test the ABSA approach based on the Bi-LSTM model. Recently, Bui Le-Minh et al. [15] proposed a “Mini Window Locating Attention” (MWLA) method which determines “mini-window” to avoid noise interference aspects and focus on sentiment-expressing words of the prescribed aspects that need to be classified as positive or negative.

3. Methodology

3.1. Data processing

Pre-processing data could be the most important component in most NLP systems because it helps us clean all the unimportant things from the text data. Due to the experimental datasets collected from e-commerce platforms, we observe that many informal words, such as teen code, abbreviations, and misspelt vocabularies, do not conform to the usual standard of the Vietnamese language. Therefore, we design a sequence of steps to pre-process the text data, which are shown in below:

- **Step 1**: Vietnamese is an accent language and users often have a mistake between accents in writing text such as “giá” (price) with “gía”, “thỏa mãn” (satisfy) with “thoả mãn”. Therefore, we standardize the accent mark and correct the free-style letters and acronyms in the text data.
- **Step 2**: We translate sentiment characters (e.g., :)), :D, etc,) and English words (quick, oke, thanks) to corresponding Vietnamese words based on the dictionary-based approach from the training and validation set.
- **Step 3**: We also remove the noise words, phrases, URLs, punctuation marks, and special characters in the sentence.
- **Step 4**: Finally, we elongated words (e.g. ngonnnn = ngon (delicious)), then, we lowercase all characters in the text input.

Unlike previous studies [2, 7], we do not apply the word segmentation technique in the pre-processing steps for T5-based models despite 85% Vietnamese words being composed of two and more syllables. The reason is that our approach is based on the pre-trained models, which are trained on text corpora without word segmentation.

Table 1. A schematic representation of different aspect categories the UIT_ABSA 2019 [16] dataset

Acronym	**Entity**	**Output Form**
asp#1	Quality	Chất lượng
asp#2	Location	Địa chỉ
asp#3	Price	Giá tiền
asp#4	Style&Option	Lựa chọn
asp#5	Miscellaneous	Vấn đề khác
asp#6	Ambience	Không gian
asp#7	Service	Phục vụ

3.2. Unified Seq2Seq Model

We consider the Aspect-category based Sentiment Analysis task as a text generation task and apply the sequence-to-sequence strategy to solve this task. Given an input sentence X of *n* words represented by $X_{1:n} = \{x_1, x_2, \ldots, x_n\}$, we aim to generate an output sequence of $Y_{1:m} = \{y_1, y_2, \ldots, y_m\}$ with unknown length of *m*. The output $Y_{1:m}$ contains all the desired aspect categories and corresponding sentiments. To facilitate the Seq2Seq model, it is necessary to linearize the labels to a natural language sequence $Y_{1:m}$. Because the category and sentiment classes are annotated in English format (e.g. *Food#Quality, positive*), therefore, we map two values to a natural language form based on a mapping dictionary. In order to build the mapping dictionary, we employ a simple statistical approach by applying a TF-IDF technique to extract the patterns for aspect categories on the training and development set.

Table 2. A schematic representation of different aspect categories for the restaurant domain [6]

Acronym	Enity#Attribute	Output Form	Acronym	Enity#Attribute	Output Form
asp#1	Restaurant#General	Nhà hàng nói chung	asp#7	Drinks#Quality	Chất lượng đồ uống
asp#2	Restaurant#Prices	Giá tiền nhà hàng	asp#8	Drinks#Prices	Giá tiền đồ uống
asp#3	Restaurant#Miscellaneous	Vấn đề khác	asp#9	Drinks#Style&Option	Lựa chọn đồ uống
asp#4	Food#Quality	Chất lượng đồ ăn	asp#10	Location#General	Địa chỉ
asp#5	Food#Prices	Giá tiền đồ ăn	asp#11	Ambience#General	Không gian
asp#6	Food#Style&Option	Lựa chọn đồ ăn	asp#12	Service#General	Phục vụ

The vocabulary with the highest score is used to represent the aspect category in Vietnamese words. Table 1, Table 2, and Table 3 show the schematic representation of different aspect categories for three benchmark datasets, respectively.

Table 3. A schematic representation of different aspect categories the UIT_ABSA 2019 [16] dataset

Acronym	Entity	Output Form
asp#1	Screen	Màn hình
asp#2	Camera	Máy ảnh
asp#3	Features	Tính năng
asp#4	Battery	Pin
asp#5	Performance	Hiệu suất
asp#6	Storage	Bộ nhớ
asp#7	Design	Thiết kế
asp#8	Price	Giá tiền
asp#9	Service and Accessories	Phục vụ hoặc phụ kiện
asp#10	General	Nói chung

Table 4. The general information of the three datasets, including the Restaurant [6], Phone [14], and Restaurant*[16]

Information	Dataset		
	Restaurant	Phone	Restaurant*
Number of samples	9737	11 122	7 828
Number of labels	12x3	10x3	7x5
Average Length	16.9	36.2	52.1
Maximum Length	96	250	208
Vocabulary size	16 773	8 063	7 521
Average Aspects/Samples	1.4	3.3	3.51

For the sentiment polarity labels, we convert them to corresponding natural words as follows: “*tốt*” if sentiment = “positive”, “*tệ*” if sentiment = “negative”, and “*tạm*” if sentiment = “neutral”. Besides, we represent the “very positive” and “very negative” sentiment classes, which are categorized into five levels according to the previous work [16], as “*rất tốt*” and “*rất tệ*”, respectively.

In the data analysis process, we found that users tend to express their opinions succinctly regarding aspect categories and their sentiment polarities, for example, “*không gian thoáng, thức ăn ngon, phục vụ nhiệt tình*” (The space is open, the food is delicious, the service is enthusiastic). Based on this observation, the natural language sentence for an aspect category and corresponding polarity is represented as the output format of the category combined with the sentiment class. For example, an output “Food#Quality, positive” is transformed as “chất lượng đồ ăn tốt”. If the input sequence has multiple aspect categories, we concatenate these corresponding mapping sentences with conjunction “và” (and) in the final target sequence Y. We also experimented with the use of specific polarity vocabulary for each aspect category. For example, “ngon” (tasty), “tệ” (bad) will represent a “positive” and “negative” class for the aspect categories with the “Quality” attribute (e.g. Food#Quality, Drinks#Quality). However, this representation reduces the overall performance of models because it increases the sparsity of sentiment polarity classes.

To convert the inference results of the model into label form, we separated the corresponding phrases into labels based on anchors. Then, we extracted information from each phrase containing keywords in the dictionary related to each aspect and polarity. For example, the model will generate the output as “lựa chọn đồ ăn tốt và chất lượng đồ ăn tốt” for the given input “-, *Món* *salad bò nướng đúng ngon luôn, mà có lẽ nên gọi bò salad đúng hơn í vì bò nhiều ơi là nhiều, gia vị có giấm táo chua ngọt dễ ăn cực kỳ.*". It can be easily transformed into two pairs of aspect categories, namely {Food#Style&Options, positive} and {Food#Quality, positive}, from the generated output.

In this paper, we employ two pre-trained sequence to sequence transformer architectures such as viT5 [17], and BARTPho [18] and their variants for the ACSA task. First, an input sequence X is fed to the encoder to obtain the hidden states:

$$h^{encoder} = Encoder(X)$$

At the i^{th} step of the decoder, the generated token y_i is computed based on the hidden states and the previous output $y_{0:i-1}$ to yield a representation:

$$h_i^{encoder} = Decoder(h^{encoder}, y_{0:i-1})$$

Finally, the output of the decoder is applied to a softmax function to calculate the probability distribution over the whole vocabulary for the next token. The loss function for the input-target pair is formulated as:

$$Loss = \sum_{i}^{m} \log p_{\theta}(y_i | h^{encoder}, y_{0:i-1})$$

where p_{θ} is the probability distribution of the next token yi, θ is initialized with the pre-trained parameter weights and further fine-tuned in the training process.

3.3. Unified Seq2Seq Model

In this paper, we explore the performance of our approach based on two latest pre-trained language models, including viT5 [17] and BARTPho[18]. The viT5 model is designed based on the original encoder-decoder T5 architecture [8]. There are two versions published for the research community[1] with the base and large models. Unlike previous pre-trained language models, viT5 models are applied a pre-processing step to normalize punctuation and capitalize as well as split digits. The authors [17] extract Vietnamese text data from the CC100 dataset [19] to pre-trained the models.

BARTPho [18] is a new large-scale monolingual model for Vietnamese inheriting from BART architecture [9]. BARTPho has two versions with different types of input texts, consisting of syllable level and word level. The word level models mean that the model is trained on word segmentation input. Each type of model has two architectures with a base and a large version. All BARTPho models are trained on 20 GB of uncompressed Vietnamese texts and released for free-research purposes[2].

4. Experimental Settings

4.1. Datasets

In this paper, we evaluate the performance of our approach based on three benchmark datasets for different domains, including restaurant [6], smartphone [14] and document-level restaurant [16]. The general information for all the datasets is reported in Table 4, while more detailed textual statistics are available in the corresponding original works. We use the same size for each data set (training, development, and testing) provided by previous works for a fair comparison. Details about the data size can be found in the corresponding studies of the original datasets.

4.2. Evaluation Metrics

Following the VLSP workshop [1], we report the micro Precision, Recall, and F1-score of all models over the test dataset of the restaurant domain. For the smartphone domain, we used the macro F1-score as previous studies [14, 20] to compare the results. A prediction is correct if all predicted aspect categories and corresponding sentiments are correct.

[1] https://github.com/vietai/ViT5

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

4.3. Settings

We adopt the pre-trained T5 models for the Vietnamese language (called as viT5 [17]) with the base and large versions from the Huggingface library [21]. Specifically, we use the base[3] and large[4] versions of the viT5 models. For the BART model, we employ the BARTPho model [18] for Vietnamese languages and their variants. The detail of the BARTpho model and its variants can be provided on the page[5].

For the restaurant domain, we used a maximum input sequence length of 128 for the sentence-level dataset [6], and 256 for the document-level dataset [16]. The output sequence length of the two datasets were 128 and 156 for sentence-level and document-level data, respectively. For the smartphone domain [14], we set the maximum input sequence length to 256 since the training data contains samples with a maximum length of 250 tokens, and the output sequence length is 128. We train the models for all experiments with a batch size of 16, 8, 4, depending on the available computational resources. All models are optimized using the AdamW optimizer. The learning rate is set to 3e-4 for the mT5 models with two versions and 2e-5 for the viT5 and PhoBART models across two domains. The experimental models are trained for 20 epochs on all datasets.

4.4. Compared Methods

To comprehensively evaluate the performance of our approach, we compare our model against the machine learning, deep learning, and Transformer-based models, which are described as follows:

SVM [2-3] is the top method in the VLSP Shared task 2018. The authors used the Support Vector Machine trained on extensive feature engineering, including n-grams, POS tags, and word features. This approach treats the aspect-category sentiment analysis as multiple binary classifications. However, this approach can not explore the related information between aspect categories in the sentence.

CNN [14, 6] is one of the deep learning architectures widely used in the text classification task. This architecture is designed based on multitask learning strategy where multiple output layers are added above the CNN model.

BiLSTM-CNN [14, 6] is an ensemble model based on the BiLSTM and CNN model for the ACSA task. It uses the BiLSTM model to learn the representation of words based on each word embedding as the input. Then, a CNN model with widths of filters (2,3,4) where 100 features each filter.

BERT [6] presented a multi-task architecture BERT that employs multiple output layers to represent the number of aspect categories with corresponding sentiments. They reported results on two benchmark datasets and demonstrated the effectiveness of this architecture over the single BERT model. Inspired by this research, we implement this approach based on different pre-trained BERT models, including PhoBERT [22], XLM-R [19], and XLM-Align [23].

mT5 [24] stands for a unified sequence-to-sequence model, in which this model is fine-tuned based on the pre-trained multilingual T5. mT5 is trained on a new Common Crawl-based dataset covering 101 languages based on a similar pre-trained text-to-text transformer model T5 [8]. This paper explores two versions, including the base and large mT5 models due to the limitation of computational resources. The configuration of these models is optimized based on the development set as our models.

[3] https://huggingface.co/VietAI/vit5-base

[4] https://huggingface.co/VietAI/vit5-large

[5] https://github.com/VinAIResearch/BARTpho

5. Results and Error Analysis

5.1. Results

Table 5, Table 6 and Table 7 present the results in terms of precision, recall, and F1-score on three benchmark datasets. The three first sections in each Table present the results of previous studies on the corresponding dataset. As shown in Table 5, our generative viT5 model outperforms all earlier and transformer-based approaches and establishes the new state-of-the-art results in two datasets.

Table 5. Main results for the restaurant domain. We use the results reported in SVM, CNN, BiLSTM-CNN, and PhoBERT [6]

Model	**Precision**	**Recall**	**F1-score**
SVM	59.35	60.37	59.85
CNN	68.61	64.91	66.71
BiLSTM-CNN	71.47	67.67	69.52
XLM-R	70.98	72.19	71.58
XLM-Align	72.08	72.08	72.08
PhoBERT	**76.78**	73.07	74.88
Ensemble BERTs [25]	75.98	74.59	75.28
BARTPho$_{syllable+base}$	68.34	67.48	67.90
BARTPho$_{word+base}$	69.01	66.57	67.76
BARTPho$_{syllable+large}$	68.10	69.19	68.64
BARTPho$_{word+large}$	71.25	69.57	70.40
mT5$_{base}$	68.01	68.66	68.33
viT5$_{base}$	73.60	73.79	73.69
mT5$_{large}$	70.52	71.45	70.98
viT5$_{large}$	75.73	**75.35**	**75.53**

Table 6. Main results for the smartphone domain. We use the results reported in SVM, CNN, BiLSTM-CNN architecture from the previous study [14]

Model	**Precision**	**Recall**	**F1-score**
SVM	16.09	23.35	19.69
CNN	33.34	22.92	27.16
BiLSTM-CNN	65.82	60.53	63.06
XLM-R	80.58	68.71	74.18
XLM-Align	79.84	69.24	74.16
PhoBERT	**82.72**	67.66	74.44
BERT+Processing [20]	80.21	79.46	78.76
BARTPho$_{syllable+base}$	79.41	80.36	79.88
BARTPho$_{word+base}$	78.86	77.40	78.12
BARTPho$_{syllable+large}$	80.49	81.25	80.86
BARTPho$_{word+large}$	79.26	80.62	79.93
mT5$_{base}$	78.91	78.89	78.90
viT5$_{base}$	78.90	**81.82**	80.33
mT5$_{large}$	79.46	79.62	79.54
viT5$_{large}$	81.06	81.20	**81.10**

Concretely, the T5 architecture achieved a performance of 75.53% in terms of the F1-score, which is higher than the state-of-the-art score of an ensemble of different BERT models [25], but the difference is not significant. However, it can be observed that our model relies less on computational complexity than the ensemble BERT model. Compared with other single BERT models, our generative model also gains better results than the monolingual PhoBERT, with an increment of +0.65% in terms of the F1-score. For the smartphone domain, the T5 model achieves results superior to the other baselines and previous performance in terms of the F1-score for the smartphone domain. Specifically, the T5 model gives the F1-score of 81.10%, which is higher than the previous result [20] with a large margin (+2.34%). These results indicate that the performances of our viT5 model are the new state-of-the-art scores for two benchmark datasets for the restaurant and smartphone domains. Particularly for the document-level dataset [16], our models demonstrate superior effectiveness compared to other baselines and approaches with a significant improvement in F1-score.

In comparison between generative models BART and T5 and their variants, it is obvious that there is a difference in the performance of the type of model. We can see that the models which are trained on monolingual corpora yield better results than multilingual models. For the T5 models, the viT5 models outperform the mT5 models with the same versions. The differences are 5.36% and 4.55% for the base and large versions for the restaurant domain, respectively. Similarly, for the document-level dataset, the difference between viT5 and mT5 models is +4.88% and +3.10% for the base and large versions, respectively. For the smartphone domain, the differences between the two models are +1.43% and +1.56% according to the base and large versions, respectively.

Comparing the performance of the BART models and the T5 models, it can be seen that the T5-based generative model consistently beat the BART-based models in all situations. The quality and the size of

Table 7. Main results on the document-level dataset for the restaurant domains [16]

Model	**Precision**	**Recall**	**F1-score**
SVM [16]	62.52	56.21	59.20
CNN	66.16	66.10	66.12
BiLSTM-CNN [7]	68.93	65.72	67.29
Multi-task Ensemble [7]	69.15	69.11	69.13
XLM-R	71.53	70.41	70.97
XLM-Align	71.08	71.64	71.36
PhoBERT	72.43	72.69	72.56
BARTPho$_{syllable+base}$	80.29	81.39	80.84
BARTPho$_{word+base}$	79.65	80.59	80.12
BARTPho$_{syllable+large}$	81.61	81.47	81.54
BARTPho$_{word+large}$	82.36	81.82	82.09
mT5$_{base}$	80.59	81.25	80.92
viT5$_{base}$	85.46	86.14	85.80
mT5$_{large}$	83.05	83.96	83.50
viT5$_{large}$	**86.24**	**86.96**	**86.60**

pre-trained corpora are the reasons to answer the results between the BARTpho [18], and viT5 [17]. While the BARTPho and its variants were trained on 20GB of news data, the viT5 models were trained on 71GB of Common Crawl data (a subset of CC100 corpus [19]) for Vietnamese language. Moreover, the text in the CC100 corpus is crawled in the websites with many sources and domains; therefore it might produce a diverse representation of Vietnamese words than the News corpus. Meanwhile, the datasets used in this paper were collected from e-commerce sites containing more non-standard vocabularies and sentences than news corpus.

5.2. Analysis

To explore the performances of aspect categories, Table 8 and Figure 1 show the F1 scores for each category in the smartphone and restaurant domains, respectively. As shown in Table 8, it can be seen that the viT5 model achieves better performances on most of the aspect categories except the “Screen”, “Features” and “Storage” aspects than the results of the previous work [20]. However, it is surprising that the method [20], which is based on the PhoBERT model combined with the simple pre-processing steps, can achieve the highest result of 95.06% for the “Storage” category. From the original work [14], the “Storage” aspect is labelled for 187 samples in the training set and 27 samples in the testing set. Compared with other aspect categories, the distribution of the “Storage” is limited to achieving the highest score when fine-tuning the transformer-based models due to a data imbalance problem. The authors [20] did not provide the reason why their models can gain the highest score for this aspect category. The performance of the baseline approach (BiLSTM-CNN architecture) from the original work [14] on this aspect category which was only 30.10% in F1-score. As for other categories in the smartphone domain, our model also gives superior performances to previous results. In particular, the results of some aspects (e.g., Camera, Battery, Performance) are improved up to 10% more than with the previously published work [20].

We noticed that our models still face the problem of imbalanced classes in the data for aspect categories with few training data samples, such as Screen, Storage, and Price. Moreover, the imbalance among sentiment classes for certain aspects is a major factor leading to ineffective results in some categories. In this case, the model might struggle to accurately predict the underrepresented classes, leading to lower performance in those categories. Moreover, we observed that our models are difficult to distinguish the polarity class due to the variability in how users express the sentiment for a specific aspect category.

As can be seen in Figure 1, the viT5 model can achieve good results for some aspect categories, such as Ambience#General, Food#Quality, and Serive#General for the restaurant domain. In contrast, there are some categories that have unpromising results related to Drinks and Restaurant entity, such as Drink#Style&Options. One of the reasons for the performance of these categories is the number of limited training samples. Besides, the ambiguity between similar classes such as Food#Prices, Drink#Prices, and Restaurant#Prices is a reason for the low performances of these aspect categories.

Table 8. Main results on the document-level dataset for the restaurant domains [16]

Category	**Le et al. [20]**			**Our best model (viT5)**		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Screen	82.51	82.29	**81.94**	79.39	79.55	79.47
Camera	80.64	79.56	78.96	88.81	89.12	**88.96**
Features	84.22	83.47	**83.16**	81.20	82.00	81.60
Battery	77.85	76.18	75.56	85.47	87.57	**86.51**
Performance	75.98	74.04	73.07	83.70	83.28	**83.49**
Storage	95.26	94.90	**95.06**	64.00	59.26	61.54
Design	77.04	75.87	75.58	84.91	83.42	**84.16**
Price	77.61	77.84	76.78	77.25	79.96	**78.58**
General	71.57	71.19	69.17	82.99	88.70	**85.75**
Ser&Acc	79.42	79.24	78.33	83.90	79.09	**81.42**

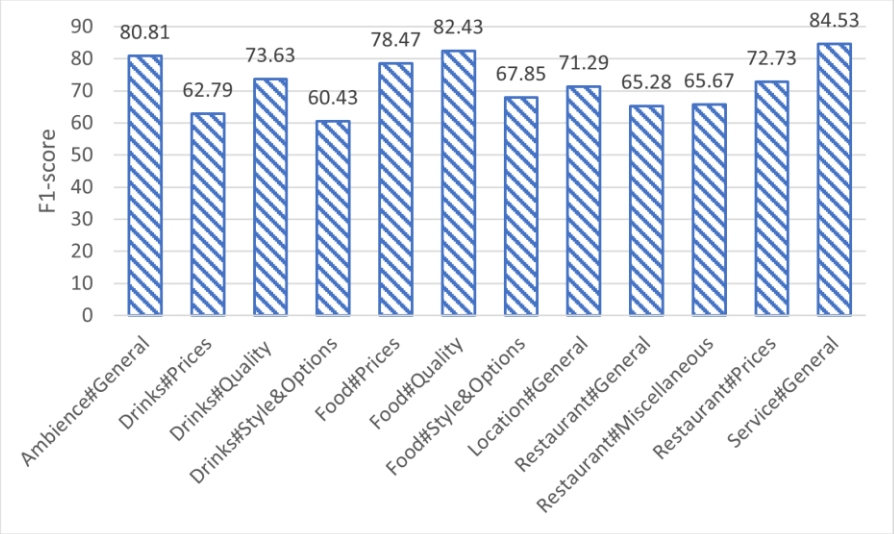

Figure 1. The F1-score of viT5 model according to 12 aspect categories for the restaurant domain.

5.3. Error Analysis

In order to deeply understand the performance of the unified text generative model, we conduct an error analysis in this section. We observed different types of errors during prediction from the viT5 model. The first type of error is that the model predicts the aspect category correctly, but the corresponding sentiment is wrong. For example, given a review in the restaurant domain “Giá thì có hơi high chút xíu.”(The price is pretty high.), the model predicted the sentiment polarity class as the “neutral” for the “Restaurant#Prices” category. This might be because the model misunderstands the “high” word in the review. This case is a type of code-mixing problem in the NLP field.

The second type of error is due to the implicit aspect category in the review. For example, the review “70K cốc trà sữa cũng đáng.” (70K for a cup of milk tea is also worth it.) is annotated as {Drinks#Quality, positive}, {Drink#Prices, neutral}. The aspect category Drinks#Quality is an implicit category because users also indirectly complement drinking water quality through the price. Therefore, our model just predicts the “Drink#Prices, neutral}” in this case.

The last type of error we noticed in the prediction is that model often gives the wrong prediction with sentences that need a lot of reasoning knowledge. For example, the sentence "Ngon hơn cả gongcha nha." (Delicious than gongcha) is labeled as the {Drinks#Quality, positive} because Gongcha is the name of a beverage restaurant. Therefore, the annotators assigned the Drink#Qualtity for the given sentence. However, the model predicted the {"Restaurant#General, positive"} for the sentence. This prediction might be true when we do not infer the Gongcha word related to drink items.

6. Conclusion and Future Work

This paper presents a transformer-based sequence-to-sequence model for the Aspect-category sentiment analysis on Vietnamese reviews. With the power of the conditional generative framework, we solve the ACSA task as the text generation problem by formatting the output labels as a natural sentence by applying a simple statistical method. We employed two famous generation models, viT5 [17], and BARTPho [18], which are pre-trained on Vietnamese corpora. Comparing the performance with previous studies and discriminative BERT-based models, our experimental results demonstrated that fine-tuning a unified generative framework viT5 achieved the best performances on three benchmark datasets than previous results and baseline models.

For future work, we can increase the performance by integrating knowledge into generation models. Internal or external knowledge can provide more information from the input and outside sources. We also plan to apply the approach in this paper for other tasks in the NLP fields.

Acknowledgments

Dang Van Thin was funded by the Master, PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), code VINIF.2022.TS120.

References

[1] H. T. Nguyen, H. V. Nguyen, Q. T. Ngo, L. X. Vu, V. M. Tran, B. X. Ngo, C. A. Le, VLSP Shared Task: Sentiment Analysis, Journal of Computer Science and Cybernetics, Vol 34, No. 4, 2018, pp. 295–310. https://doi.org/10.15625/1813-9663/34/4/13160.

[2] D. V. Thin, V. D. Nguyen, N. V. Kiet, N. L. T. Ngan, A Transformation Method for Aspect-based Sentiment Analysis, Journal of Computer Science and Cybernetics, Vol 34, No. 4, 2018, pp. 323–333. https://doi.org/10.15625/1813-9663/34/4/13162.

[3] N. T. T. Thuy, N. X. Bach, T. M. Phuong, Leveraging Foreign Language Labeled Data for Aspect-based Opinion Mining, International Conference on Computing and Communication Technologies, 2020, pp. 1–6.

[4] M. Schmitt, S. Steinheber, K. Schreiber, B. Roth, Joint Aspect and Polarity Classification for Aspect-based Sentiment Analysis with End-to-End Neural Networks, Conference on Empirical Methods in Natural Language Processing, 2018, pp.1109-1114.

[5] H. Cai, Y. Tu, X. Zhou, J. Yu, R. Xia, Aspect-Category based Sentiment Analysis with Hierarchical Graph Convolutional Network, 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, 2020, pp. 833–843.

[6] V. D. Thin, N. L.-T. Nguyen, T. M. Truong, L. S. Le, D. T. Vo, Two New Large Corpora for Vietnamese Aspect-based Sentiment Analysis at Sentence Level, Transactions on Asian and Low-Resource Language Information Processing, Vol 20, No. 4, 2021, pp 1–22. https://doi.org/10.1145/3446678.

[7] D. V. Thin, L. S. Le, H. M. Nguyen, N. L.-T. Nguyen, A Joint Multi-task Architecture for Document-level Aspect-based Sentiment Analysis in Vietnamese, International Journal of Machine Learning and Computing, Vol 12, No. 4, 2022, pp 126-135. DOI: 10.18178/ijmlc.2022.12.4.1091.

[8] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, Journal of Machine Learning Research. Vol 21, No. 140, 2020, pp 1–67.

[9] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7871–7880.

[10] J. Liu, Z. Teng, L. Cui, H. Liu, Y. Zhang, Solving Aspect Category Sentiment Analysis as a Text Generation Task, Conference on Empirical Methods in Natural Language Processing, 2021, pp. 4406–4416.

[11] W. Zhang, Y. Deng, X. Li, Y. Yuan, L. Bing, W. Lam, Aspect Sentiment Quad Prediction as Paraphrase Generation, Conference on Empirical Methods in Natural Language Processing, 2021, pp. 9209–9219.

[12] W. Zhang, X. Li, Y. Deng, L. Bing, W. Lam, Towards Generative Aspect-Based Sentiment Analysis, 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021, pp. 504–510.

[13] D. V. Thin, V. D. Nguyen, K. V. Nguyen, N. L.-T. Nguyen, Deep Learning for Aspect Detection on Vietnamese Reviews, 5th NAFOSTED Conference on Information and Computer Science, 2018, pp. 104–109.

[14] L. P. Luc, P. H. Pham, K. T. T. Nguyen, S. K. Huynh, T. T. Nguyen, L. T. Nguyen, T. V. Huynh, K. V. Nguyen, SA2SL: From Aspect-Based Sentiment Analysis to Social Listening System for Business Intelligence, International Conference on Knowledge Science, Engineering and Management, 2021, pp. 647–658.

[15] B. L. Minh, T. P. Le, K. H. Tran, K. H. Bui, H. Q. Le, D. C. Can, H. N. C. Thanh, M. V. Tran, Aspect-Based Sentiment Analysis Using Mini-Window Locating Attention for Vietnamese E-commerce Reviews, 13th International Conference on Knowledge and Systems Engineering, 2021, pp. 1-4.

[16] M. H. Nguyen, T. M. Nguyen, D. V. Thin, N. L. T. Nguyen, A Corpus for Aspect-based Sentiment Analysis in Vietnamese, 11th International Conference on Knowledge and Systems Engineering, 2019, pp. 1–5.

[17] L. Phan, H. Tran, H. Nguyen, T. H. Trinh, ViT5: Pretrained Text-to-Text Transformer for Vietnamese Language Generation, Conference of the North American Chapter of the Association for Computational Linguistics, 2022, pp. 136–142.

[18] N. L. Tran, D. M. Le, D. Q. Nguyen, BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese, 23rd Annual Conference of the International Speech Communication Association, 2022.

[19] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, É. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451.

[20] B. H. Le, H. M. Nguyen, N. K. P. Nguyen, B. T. Nguyen, A New Approach for Vietnamese Aspect-Based Sentiment Analysis, 14th International Conference on Knowledge and Systems Engineering, 2022, pp. 1–6.

[21] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-Art Natural Language Processing, Conference on Empirical Methods in Natural Language Processing, 2020, pp. 38–45.

[22] D. Q. Nguyen, A. T. Nguyen, PhoBERT: Pre-trained language models for Vietnamese, Findings of the Association for Computational Linguistics: EMNLP, 2020, pp. 1037–1042.

[23] Z. Chi, L. Dong, B. Zheng, S. Huang, X.-L. Mao, H.-Y. Huang, F. Wei, Improving Pretrained Cross-Lingual Language Models via Self-Labeled Word Alignment, 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021, pp. 3418–3430.

[24] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer, Conference of the North American Chapter of the Association for Computational Linguistics, 2021, pp. 483–498.

[25] V. D. Thin, D. N. Hao, N. L.-T. Nguyen, An Effective Contextual Language Ensemble Model for Vietnamese Aspect-based Sentiment Analysis, 9th NAFOSTED Conference on Information and Computer Science, 2022, pp. 30–34.