A Systematic Literature Review on Vietnamese Aspect-based Sentiment Analysis

DANG VAN THIN, DUONG NGOC HAO, and NGAN LUU-THUY NGUYEN, University of Information Technology, Ho Chi Minh City, Vietnam and Vietnam National University, Ho Chi Minh City

Aspect-based sentiment analysis (ABSA) is one of the principal tasks in the automatic deep understanding of texts, widely applied in a broad range of real-world applications. Many studies have been performed on different tasks and datasets for other languages (e.g., English, Chinese) to address this topic. For Vietnamese language, this topic has been attracting considerable interest in recent years. However, we found that many studies tend to repeat the research instead of inheriting and extending the previous works. Moreover, previous studies' methods of comparison or evaluation metrics have not shown consistency and connection. This might restrict the development of future studies on this research topic. To the best of our knowledge, no research has been conducted to overview the existing studies for the ABSA research in Vietnamese language. The primary objective of this study is to provide a systematic and comprehensive review of the current Vietnamese ABSA research. More specifically, we analyze the early approaches, evaluation metrics, and available published benchmark datasets used in the Vietnamese ABSA task. We also discuss the challenge and recommend potential future directions for Vietnamese ABSA. This work is expected to provide readers with a wealth of knowledge, the research gap, and the challenges in the Vietnamese ABSA field.

CCS Concepts: • **Computing methodologies → Natural language processing;**

Additional Key Words and Phrases: Aspect-based sentiment analysis, systematic literature review, methods, datasets, Vietnamese language

ACM Reference format:
Dang Van Thin, Duong Ngoc Hao, and Ngan Luu-Thuy Nguyen. 2023. A Systematic Literature Review on Vietnamese Aspect-based Sentiment Analysis. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 22, 8, Article 218 (August 2023), 28 pages.
https://doi.org/10.1145/3610226

1 INTRODUCTION

With the development of e-commerce platforms and online shopping needs, a large amount of user-generated content is increasing as users share their opinions online, especially during the COVID-19 pandemic. These data are very important to the growth of service or product providers as well as new users. For example, business organizations can improve their products based on

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

Authors' address: D. Van Thin, D. Ngoc Hao, and N. Luu-Thuy Nguyen (corresponding author), University of Information Technology, Ho Chi Minh City, Vietnam, and Vietnam National University, Ho Chi Minh City, Vietnam; emails: thindv@uit.edu.vn, haond@uit.edu.vn, ngannlt@uit.edu.vn.

2375-4699/2023/08-ART218 $15.00
https://doi.org/10.1145/3610226

what consumers want via user reviews. Sentiment analysis is one of the principal tasks in the automatic deep understanding of texts, widely applied in a broad range of real-world applications. There are three levels of sentiment analysis research, including document level, sentence level, and aspect level [20]. With the first two levels, the aim of this task is to identify the overall sentiment classes for the document and sentence review. However, with the demand for deep analysis in review analytics, the **Aspect-Based Sentiment Analysis (ABSA)** has received increasing attention in recent years [4, 26, 56]. The purpose of this task is to detect the sentiment polarity of a certain aspect instead of an entire sentence or document review. Therefore, the ABSA research provides more useful information for downstream applications.

In general, the ABSA topic consists of many sub-tasks related to four aspect-based sentiment elements, including aspect categories, aspect terms, opinion terms, and sentiment polarity [55]. The elements can be studied individually or combined with other elements to form more difficult tasks. For example, the Aspect category detection is a single task that aims to extract the list of categories mentioned in the review. While the Aspect-Category Sentiment Analysis is one of the compound tasks that extract the pair of aspect categories and their corresponding sentiments. For that reason, the ABSA tasks have been studied in many languages and presented through many survey papers such as English [4, 26, 56], Arabic [3, 33], and so on. For low-resource languages such as Vietnamese, there are some existing studies focusing on the ABSA tasks, including publishing benchmark datasets [29, 30, 40, 45, 48, 51] and presenting the approaches [24, 41, 42, 44, 46]. However, we found that many studies tend to repeat the previous studies instead of inheriting and extending the previous works. In addition, no research has conducted a detailed review of the existing Vietnamese ABSA studies in a systematic way to show what has been done and recommend future directions.

Hence, the major objective of this study is to provide a systematic and comprehensive review of the current Vietnamese ABSA research. In this article, different approaches and evaluation metrics used in the ABSA task are analyzed and discussed to explore the advantages and disadvantages. Moreover, the available published benchmark datasets for Vietnamese ABSA are provided with detailed information for the community. This work is presented with the goal of systematizing the development of ABSA research in the Vietnamese language. This work is highly beneficial to provide readers with a wealth of knowledge, the research gaps, and the challenges in the Vietnamese ABSA. The main contributions can be summarized as follows:

- Several literature manuscripts have been analyzed to identify existing methods and approaches from previous studies to keep up with the current trending research.
- Surveying and analyzing the published benchmark Vietnamese ABSA datasets for further studies to inherit and develop a variety of different tasks.
- Presenting the challenges, research gaps, and future research directions for Vietnamese ABSA research.

The remainder of this article is structured as follows: Section 2 describes a background of aspect-based sentiment analysis and the Vietnamese language. Section 3 presents the research methodology for a systematic literature review. Section 4 provides a literature review of the current Vietnamese ABSA studies. Section 5 discusses the main findings, challenges, and future research directions. The conclusion is presented in Section 6.

2 CONCEPTUAL BACKGROUND

2.1 Aspect-based Sentiment Analysis

ABSA—a subfield of Sentiment Analysis that has attracted the attention of diverse research groups. The ABSA is defined as a problem to identify a sentiment expressed towards a specific aspect in the

Table 1. An Illustrative Example of the ABSA Tasks

Input Review: S = The food is delicious, but the staff is unfriendly.		
Type	**Task**	**Output**
Single Task	Aspect Term Extract (ATE)	+ food + staff
	Aspect Category Detection (ACD)	+ food#quality + service#general
	Opinion Term Extraction (OTE)	+ delicious + staff
	Aspect Sentiment Classification (ASC)	+ food → positive + staff → negative

Compound Task	Aspect-Opinion Pair Extraction (AOPE)	+ {food,delicious}
		+ {staff,unfriendly}

	Aspect-Term Sentiment Analysis (ATSA)	+ {food,positive}
		+ {staff,negative}

	Aspect-Category Sentiment Analysis (ACSA)	+ {food#quality,positive}
		+ {service#general,negative}

	Aspect Sentiment Triplet Extraction (ASTE)	+ {food,delicious,positive}
		+ {staff,unfriendly,negative}

	Aspect Category Sentiment Detection (ACSD)	+ {food,food#quality,positive}
		+ {staff,service#general,negative}

	Aspect Sentiment Quad Prediction (ASQP)	+ {food,food#quality,delicious,positive}
		+ {staff,service#general,unfriendly,negative}

text review [34]. According to the work of Liu [20], most tasks in the ABSA research are developed based on five elements of the quintuple $\{a_i, c_{ij}, o_{ijkl}, p_k, t_l\}$, where a_i is an entity or aspect term, c_{ij} is the jth category of a_i, while o_{ijkl}, p_k, t_l are the opinion holder, sentiment class, and time, respectively. Different ABSA tasks are derived from the above quintuple, including single tasks and compound tasks. Table 1 shows an example of different current tasks in ABSA research. The definitions of tasks are presented as follows:

- **Aspect Term Extraction (ATE):** Given a review S with n words $\{X = w_1, w_2, \ldots, w_n\}$, the ATE task aims to extract explicit m aspect terms $\{Y = a_1, a_2, \ldots, a_m\}$, where a_i is a subsequence of the input review S.
- **Aspect Category Detection (ACD):** Given a review S consisting of n words $\{X = w_1, w_2, \ldots, w_n\}$, a pre-defined set C of aspect categories, the ACD task aims to detect the list of m aspect categories mentioned in the review $\{Y = c_1, c_2, \ldots, c_m\}$, where c is an aspect category in C.
- **Opinion Term Extract (OTE):** Given a review S with n words $\{X = w_1, w_2, \ldots, w_n\}$ and an aspect a, the OTE task aims to extract explicit opinion expression toward an aspect, where opinion expression starts from position i to position j with $0 \leq i \leq j \leq n$.
- **Aspect Sentiment Classification (ASC):** Given a review S with n words $\{X = w_1, w_2, \ldots, w_n\}$ and a specific aspect term or aspect category, the ASC task aims to predict the polarity class p_i toward aspect term or category, where p_i belongs to a pre-defined set P of sentiment polarities.
- **Aspect-Opinion Pair Extraction (AOPE):** Given a review S with n words $\{X = w_1, w_2, \ldots, w_n\}$, the purpose of this task is to extract the pairs of m aspect-opinion expression $\{Y = \{a_1; o_1\}, \ldots, \{a_m; o_m\}\}$, where o_i is an opinion expression for the aspect term a_i.
- **Aspect-Term Sentiment Analysis (ATSA):** Given a review S with n words $\{X = w_1, w_2, \ldots, w_n\}$ and a pre-defined set P of sentiment polarities, the purpose of this task is to extract the pairs of m aspect-opinion expressions $\{Y = \{a_1; p_1\}, \ldots, \{a_m; p_m\}\}$, where p_i is a sentiment polarity classes for the aspect expression a_i.

- **Aspect-Category Sentiment Analysis (ACSA):** Given a review S consisting of n words $\{X = w_1, w_2, \ldots, w_n\}$, a pre-defined set C of aspect categories, and a pre-defined set P of sentiment polarities, the purpose of ACSA is to identify all pairs (*c,p*), where c is a category in C and p is a sentiment class in P.
- **Aspect Sentiment Triplet Extraction (ASTE):** Given a review S with n words $\{X = w_1, w_2, \ldots, w_n\}$, the output of the ATSA task is the list of the m triplet information from a given input $\{Y = \{a_1; o_1; p_1\}, \ldots, \{a_m; o_m; p_m\}\}$, where aspect term a_i is expressed though opinion expression o_i with the p_i sentiment class.
- **Aspect Category Sentiment Detection (ACSD):** Given a review S with n words $\{X = w_1, w_2, \ldots, w_n\}$, a pre-defined set C of aspect categories, and a pre-defined set P of sentiment polarities, the ACSD task is aim to extract the list of the m triplets $\{Y = \{a_1; c_1; p_1\}, \ldots, \{a_m; c_m; p_m\}\}$, where aspect term a_i is the expression of category c_i, p_i sentiment class of aspect term a_i, and category c_i.
- **Aspect Sentiment Quad Prediction (ASQP):** This task is a complete task in the ABSA research with the purpose of predicting all the four sentiment elements $\{a_i, c_i, o_i, p_i\}$ in the input. Given a review S with n words $\{X = w_1, w_2, \ldots, w_n\}$, a pre-defined set C of aspect categories, and a pre-defined set P of sentiment polarities, the ASQP is to extract the list of m quadruplet form m triplets $\{Y = \{a_1; c_1; o_1; p_1\}, \ldots, \{a_m; c_m; o_m; p_m\}\}$, where a_i is an aspect term for the ith category c and has the p_i polarity class with the corresponding o_i opinion expression.

2.2 Vietnamese Language

Vietnamese is spoken by the largest population and is the official language of Vietnam. The linguistics research classified the Vietnamese language in the Mon-Khmer branch of the Austroasiatic family [1]. It is also one of the few languages in Asia that uses the Latin alphabet rather than other symbols. We present some basic information related to the language type and vocabulary characteristics of Vietnamese are presented below:

- **Alphabet:** The Vietnamese Alphabet consists of 12 vowel letters, 17 consonant letters, 11 consonant letter clusters, 27 diphthong letters, and 17 triphthong letters. Numerous Vietnamese vocabulary was borrowed numerous from Sino-Vietnamese (e.g., đại họ (university)) and French (e.g., cà-phê (café)). In addition, modern Vietnamese words have been borrowed from other languages such as English (e.g., tivi (television)), Japanese (e.g., su-mô (sumo)), and so on.
- **Dialectical:** Vietnam country consists of three main geographical regions (northern, central, and southern). Therefore, there are many variations in the lexicon, and they depend on the region. For example, to describe the word "we," the dialect used in northern, central, and southern is "chúng tôi," "bọn tui," and "tụi tui," respectively.
- **Monosyllabic:** Each word in Vietnamese has a syllable; however, two or three syllables can be composed to create a new vocabulary. For example, hạnh phúc (happy), vui vẻ (funny), tuyệt vời (perfect).
- **Non-inflectional:** It means that Vietnamese vocabulary does not change forms to express grammatical categories, e.g., tense, case. Moreover, no third-person or singular and plural agreement in Vietnamese words. Therefore, the meanings depend on the context of the sentence and the quantifiers, time words. In addition, a single word is difficult to identify the **Part-of-Speech (PoS)** without existing in a sentence.
- **Tonal:** Vietnamese is a tonal language with six different tones represented by five tonal accent marks. Using the wrong tonal can significantly change the meaning of a word in Vietnamese.

Table 2. Research Questions for Our Systematic Literature Review

Research Question	Motivation
RQ1. What are the different tasks of aspect-based sentiment analysis in Vietnamese?	To explore the current tasks in different domains of Vietnamese ABSA research.
RQ2. Which approaches have been used in different primary studies?	To investigate the type of approaches, methods, and algorithms used in different tasks of Vietnamese ABSA research.
RQ3. What evaluation metrics are used to measure the performance of Vietnamese ABSA tasks?	To identify metrics used for different ABSA tasks.
RQ4. Which datasets have been published for free-research purposes?	To explore the information of annotated datasets for different ABSA tasks in Vietnamese, which are published for the research community.
RQ5. What are the challenges in the Vietnamese ABSA task?	To indicate the challenge in ABSA research related to the linguistic characteristics of Vietnamese and domains of user review.
RQ6. What are the research gaps in the current Vietnamese ABSA studies?	To show the research gaps and future directions in ABSA studies.

- **Pronoun:** Vietnamese is one of the languages with a rich system of pronouns. It depends on many factors when used, such as time, context, relationship, social position, and so on.
- **Grammar:** The grammatical relations depend on word order and tool words in the sentence.

3 RESEARCH METHODOLOGY

To tackle the research questions in this article, we applied the following steps for a systematic search [14], including (1) developing a review Protocol to outline the structure of keywords; (2) conducting a search strategy from data sources based on the titles, keywords, and abstracts to collect the relevant papers; (3) extracting the content to provide a thorough description, results, and identifying the gaps for ABSA tasks in the Vietnamese language from the selected papers.

- **Protocol Development:** This component is developing to improve the quality of systematic literature review (e.g., reducing the possibility of research bias, main research questions). In this step, the main research questions, which are presented in Table 2, were established to address in this article.

  These research questions were planned to decide how we do with the next search strategy by identifying the sets of keywords and data sources.
- **Data Sources and Search Strategy:** It is necessary to determine search strings. We designed two sets of keywords and combined them with wildcard symbols and Boolean operators to enhance the search's quality. We search the keywords on popular search engineers such as Google Scholar, as well as several other electronic databases such as Scopus, **Asian Citation Index (ACI)**, ACM Digital Library, IEEE Xplore Digital Library, SpringerLink Digital Library, ScienceDirect Digital Library, arXiv by Cornell University, World Scientific, and so on. Especially, we also seek keywords in Vietnamese at Vietnam National University Library to collect the papers in national conferences or journals. This process helps us to ensure that relevant documents are extracted. Then, we conduct a double-check step through the papers independently based on several criteria to make a valuable contribution to the review. Finally, 31 research publications were selected as offering relevance for this study.
- **Data Extraction from the Selected Publications:** The objective of this stage aims to collect all the information needed to address the research questions. We extract the following primary information from each review paper: (1) The bibliographic data of the papers; (2) The task coverage in the study; (3) The main contributions of the research paper; (4) The findings of the research paper. Table 3 presents the details of information extraction and

Table 3. Data Extraction Description

Information Extraction	Descriptions
1. Bibliographic data	Title, Author, Year, Type of Manuscript
2. Task Coverage	Which is the primary task the focus of the study?
3. Approaches, Algorithms	What type of approaches, methods, and algorithms have been used in the study?
4. Validation strategy	What validation strategy has been used to evaluate the model?
5. Comparative methods	Which previous methods, approaches, and algorithms are compared in the study?
6. Evaluation metrics	What type of metrics have been used to evaluate the performance of models?
7. Dataset	What dataset has been used by the study, including the size and domain of data?
8. Experimental result	What are the results of the test data in the study?

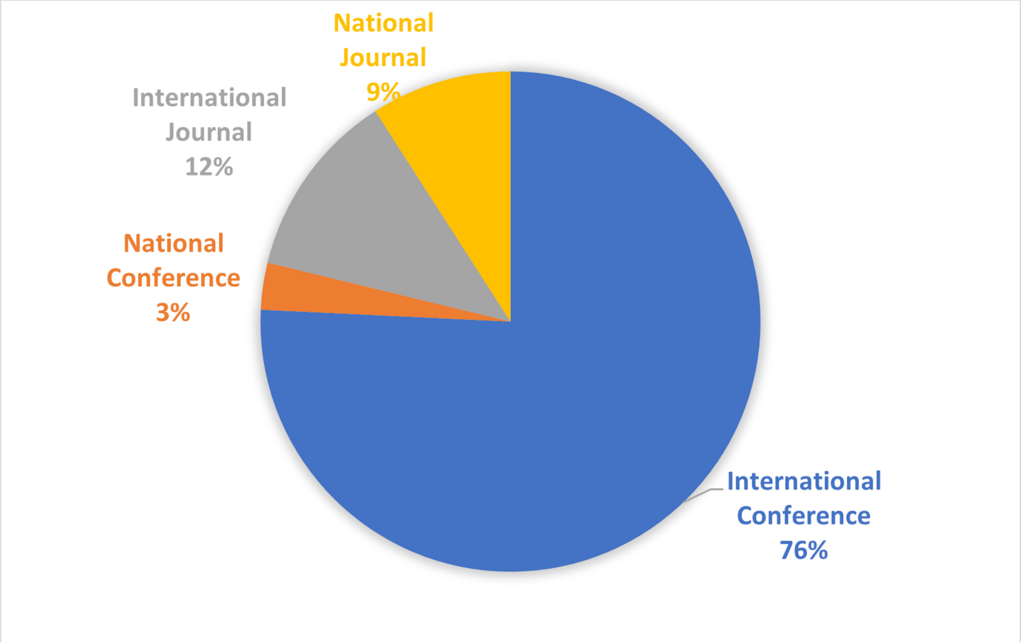

Fig. 1. Distribution of selected papers according to their types.

its descriptions extracted from each surveyed publication. We selected the papers from the international conference/journal as well as the national conference/journal (see Figure 1).

4 LITERATURE REVIEW

This section provides answers to the research questions presented in Section 3. To the authors' knowledge, there are only 31 manuscripts about the ABSA research field in Vietnamese language up to now, where 7 were published in journals, and 24 papers were published in conferences or workshops. Figure 2 shows the number of surveyed papers during different years. First, we give the summary of the current task coverage of ABSA research in Vietnamese language (answered

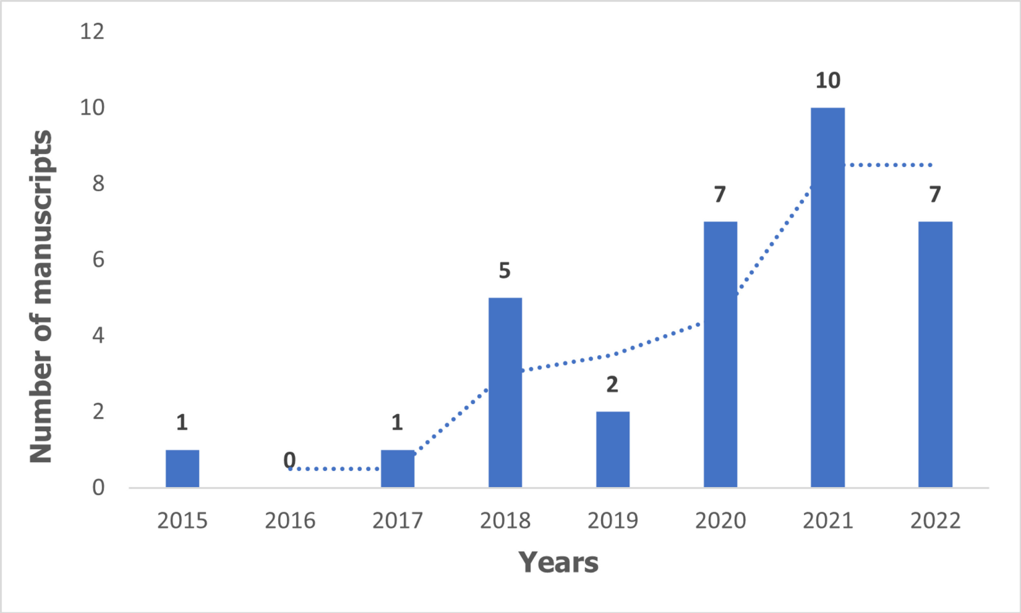

Fig. 2. Publication trend for Vietnamese aspect-based sentiment analysis.

for RQ1). Second, we present the detail of approaches, methods, and algorithms of previous studies (answered for RQ2). Third, we summarize the information of available benchmark datasets for Vietnamese language (answered for RQ4). Finally, we report the evaluation metrics for each specific task (answered for RQ3).

4.1 Task Coverage

Unlike other resource-rich languages like English [56], which have received increasing attention in the past decade, the studies in the Vietnamese language focus on some specific sub-tasks in the ABSA research. In this section, we present the task definitions and their examples, which are studies by previous research works in the Vietnamese community. Figure 3 shows the distribution of the 31 manuscripts according to the task coverage. Besides the tasks presented in Section 2.1, the work of Thanh et al. [40] proposed a new task called span detection task for the Vietnamese ABSA topic that is formulated as below:

- **Span Detection:** Given a document-level review S with words for a specific domain, the aim of this task is to extract one or more spans of opinions for each aspect category, where each span starts from position i to position j with $0 \leq i \leq j \leq n$.

As shown in Figure 3, it can be seen that most previous studies focused on the Aspect Category Detection task and Aspect-Category Sentiment Analysis task on the ABSA topics. The primary reason to answer this situation is that there are publicly datasets only annotated for two tasks. Moreover, the sizes of these available datasets are large; therefore, they are used by the research community in Vietnamese.

4.2 Approaches

To make it consistent with the previous section, we present the overview of the previous approaches grouped by the task coverage through the corresponding paragraphs. It means that the first paragraph presents the approaches for the Aspect Category Detection task.

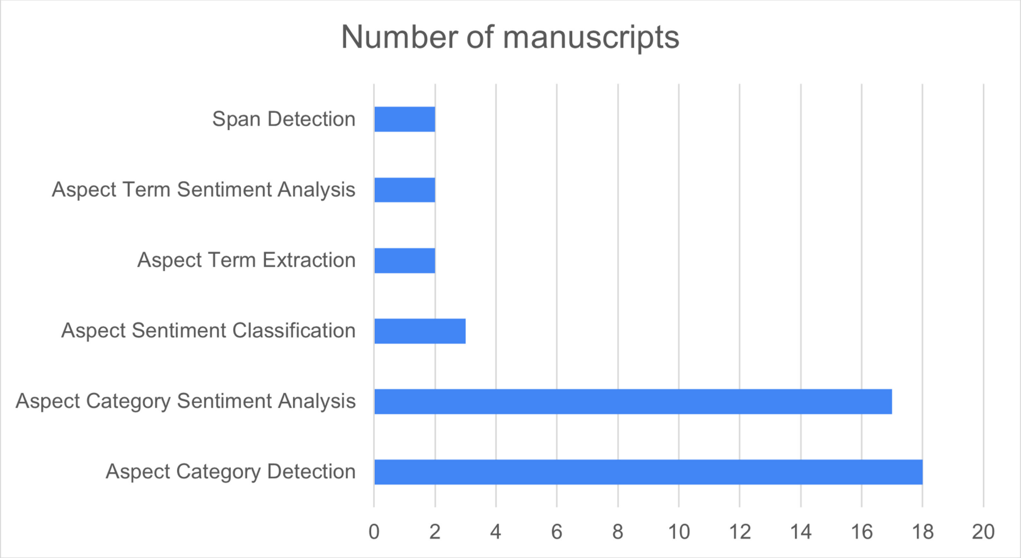

Fig. 3. Distribution of 31 selected papers according to the task coverage.

4.2.1 *Aspect Category Detection.*

The brief information of selected papers for the Aspect Category Detection task is shown in Tables 4 and 5.

First, Thin et al. [46] proposed a transformation approach by treating the ACD task as a multiple binary classification problem. The authors trained binary classifiers corresponding to the number of aspect categories for a specific domain based on linear SVM models with handcraft features. The TF-IDF technique uses the list of handcraft features to convert them to the vector. Besides, the authors also applied the list of pre-processing steps to normalize the input review. This approach achieved the best performance in two document-level datasets compared with other participants in the shared-task VLSP in 2018.

As similar, the work of Thuy et al. [47] and Nguyen et al. [30] also employed this approach to solve the ACD task on the sentence-level dataset. The feature engineering technique is the main difference between the three previous works (see Table 4). However, we found that this approach encounters problems with aspect categories of fewer training samples and imbalanced classes in the training process. To solve this challenge, Thuy et al. [47] utilizes existing data from another language to improve the overall performance, while Nguyen et al. [30] applied the SMOTE technique to balance training samples between classes. We can also observe that data expansion by translating from other languages will depend on the quality of machine translation algorithms. Moreover, the style of review between Vietnamese and other languages is quite divergent; therefore, the difference in performance is not significant. We can also observe that the transformation approach cannot utilize the correlation between aspect categories because it trains each category independently.

Instead of solving the ACD as the binary classification problem, Reference [8] used the multi-label representation to encode the output for aspect categories. The authors used the one-hot encoding and the Chi-square method to represent the input. However, the one-hot representation will increase the sparsity of the feature representation and cannot yield good results in this case. Moreover, the authors applied three resampling techniques (Oversampling, Undersampling, and SMOTE) to tackle the imbalance problem in the dataset. The experiments used the SVM, Naive Bayes, Random Forrest, and Logistic Regression as the learning models. Then, a bagging

Table 4. Summary of Machine Learning Approaches for the Aspect Category Detection Task

Papers	Method	Pre-processing steps	Datasets	Results	Advantages	Disadvantages
Thin et al. [46]	+Multiple SVM classifiers with handcraft features Binary classification problem	+Word normalization +Removing special characters, words +Word standardization +Correcting elongated words +Word segmentation	VLSP2018	+Restaurant: 77% +Hotel: 69%	+Easy to implement approach +Applying various processing steps +Using different handcraft features	+Cannot utilize the correlation between categories +Imbalance problem during training +Depend on the feature engineer
Thuy et al. [47]	+Data augmentation by translating the from other languages. + Linear SVM classifiers with handcraft features	N/A	Self-labelling	+Restaurant:76.62%	+Utilizing existing data from another language to improve the performance +Designing variety of handcraft features	+Cannot utilize the correlation between categories +Imbalance problem during training +Depend on the feature engineer
Thin et al. [42]	+Deep Convolutional Neural Network +Multi-label classification	+Word normalization +Removing special characters, words +Word standardization +Correcting elongated words +Word segmentation	VLSP2018	+Restaurant: 80.40% +Hotel: 69.25%	+Applying various processing steps +No require feature engineer	+No compare with other deep learning models +Require a large annotated dataset +Computational complexity
Nguyen et al. [30]	+Multiple Linear SVMs classifiers +Multi-class classification	+Replace special words +Lowercase sentence +Word segmentation	Self-labelling	+Restaurant: 87.13%	+Easy to implement approach +Applying the SMOTE technique	+Cannot utilize the correlation between categories +Require feature engineer +Depend on the feature engineer
Bui and Nguyen [5]	+Ensemble deep learning approach +Convolution Neural Network +Long short-term Memory	+Stopword removal +Noise removal +Word segmentation	VLSP2018	+Restaurant: 79.50% +Hotel: 69.60%	+Easy to implement approach +No feature engineer component	+No comparison with previous studies +Combine the advantage of two deep learning models +Require a large annotated dataset +Computational complexity

* We denote the "self-labeling" value to indicate the datasets provided in the same studies.

Table 5. Summary of Machine Learning Approaches for the Aspect Category Detection Task (Continued)

Papers	Method	Pre-processing steps	Datasets	Results	Advantages	Disadvantages
Nguyen et al. [32]	+Traditional machine learning algorithms +Multi-label classification +One-hot vs Chi-square encoding +Resampling methods	+Word segmentation +Remove repeated characters +Delete word with length > 7 +Stopword removal +Replace acronyms +Remove non-word characters	Self-labeling	86.45%	+Easy to implement approach +Applying various processing steps +Using different resampling method to handle the imbalance problem	+Cannot utilize the correlation between categories +Depend on the feature engineer
Cong et al. [8]	+Deep learning architectures. +Text representation techniques +Caculate the importance's weight of words	+Word segmentation +Stopword removal +Remove punctuation +Remove duplicate words +Remove nonsense words +Replace acronyms +Remove non-word characters	Self-labelling	+	+Propose a method to calculate the weight of words + Explore the performance of text representation methods	+Cannot utilize the correlation between categories +Imbalance problem during training +Computational complexity
Van Thin et al. [53]	+BERT-based learning approach. +Multi-label Classification +Multilingual and monolingual BERTology models	+Word normalization +Removing special characters, words +Word standardization +Correcting elongated words +Word segmentation	UIT_ABSA	+Restaurant: 86.53% +Hotel: 79.16%	+Different pre-trained language models are investigated +Applying various processing steps	+Require a large annotated dataset +Computational complexity

* We denote the "self-labeling" value to indicate the datasets provided in the same studies.

ensemble architecture was employed on four based classifiers. Unfortunately, the authors did not compare the results with the deep learning or transformer-based model, because these models are effective for this type of multi-label classification problem.

With the power of deep learning models, Thin et al. [42] presented a deep convolution neural network based on four convolutional blocks to capture more information between aspect categories. Then, the authors applied a threshold to assign category classes for the input review. Compared to the transformation approach [46], this model produced better performance on two VLSP2018 benchmark datasets. The limitation of this work is that the authors do not compare the proposed method with other deep learning models such as CNN or LSTM, and so on.

The work of Bui and Nguyen [5] explored the performance of CNN, LSTM, and CNN-LSTM models on two VLSP2018 datasets for the ACD task. Their CNN-LSTM architecture is designed by taking advantage of the CNN model to extract the features, then these features are fed directly to the LSTM model to predict the outputs. The experimental results have shown that the combination model CNN-LSTM is able to achieve better results in terms of F1 than single models. Although the authors conducted the experiments on the same datasets as the previous studies [42, 46], the authors did not compare them with SOTA results to demonstrate the effectiveness of the proposed models. We found that there is another work [12] by the same author presenting similar contributions to this work [5]. The difference between the two works is that one is written in English, and the other is written in Vietnamese. Unfortunately, this is related to the problem of self-plagiarism in scientific research.

Instead of relying entirely on the deep learning model, Nguyen et al. [32] investigated different representation methods, including one-hot vector, static word embedding, and contextual word embedding combined with two deep learning models (CNN and MLP). Besides, the author proposed three methods to learn the weights of essential words in the input, including the Chi-square, Self-attention, and Micro-context. The CNN model with the Fasttext word embedding, Micro-context attention, and Chi-square achieved the best overall performance.

Recently, the development of BERT has brought a new revolution in the NLP field. As a result, Van Thin et al. [53] presented an evaluation work to investigate the performance of different pre-trained language models on two sentence-level benchmark datasets. The authors designed the based model based on the original work of the BERT model [10] by fine-tuning the pre-trained language models on the ACD task. They conduct the experiments on three monolingual and three multilingual models where models are trained on the corpus with different domains, sizes, and configurations. The purpose of this article is aim to explore where is the best pre-trained language models for fine-tuning the ACD task in Vietnamese. The experimental results demonstrated that fine-tuning the PhoBERT model [28] achieved the best performances on two datasets. One of the reasons for the best performance of PhoBERT is the training corpus that is applied to the word segmentation technique in the pre-processing process. Noticing that most Vietnamese words are composed of at least two or more syllables [13]. Moreover, the authors also conducted a small experiment by combining different training data from other languages and utilized the multilingual language model to improve the overall performance. This is an interesting point in case there are not enough resources to annotate a large dataset.

4.2.2 Aspect-Category Sentiment Analysis. Aspect-Category Sentiment Analysis is one of the active tasks engaging the community in Vietnamese ABSA. The purpose of this task is aim to detect the aspect categories and corresponding sentiment polarities simultaneously. Table 6 and Table 7 show our summarization of the manuscript for the ACSA task. To address this task, References [46, 48] presented a straightforward approach by first identifying the list mentioned categories in the review, then predicting the sentiment class for each detected aspect category (see Figure 4(1)).

Table 6. Summary of Machine Learning Approaches for the Aspect-Category Sentiment Analysis Task

Papers	Method	Pre-processing steps	Datasets	Results	Advantages	Disadvantages
Thin et al. [46]	+Two-stage approach based on ML algorithms. +Multiple SVM classifiers with handcraft features	+Word normalization +Removing special characters, words +Word standardization +Correcting elongated words +Word segmentation	VLSP2018	+Restaurant: 77% +Hotel: 69%	+Easy to implement approach +Applying various processing steps +Using different handcraft features	+Cannot utilize the correlation between categories +Imbalance problem during training +Depend on the feature engineer
Nguyen et al. [30]	+Multiple SVM classifiers with handcraft features Feature augmentation using SMOTE technique +Multi-task learning between aspect category with sentiment polarity	+Replace special words +Lowercase sentence +Word segmentation	Self-labeling	+Restaurant: 59.20%	+Easy to implement approach +Using different handcraft features	+Cannot utilize the correlation between categories +Imbalance problem during training +Depend on the feature selection approach
Thuy et al. [48]	+Two-stage approach based on ML algorithms. +Multiple SVM classifiers with handcraft features Data augmentation using back-translation	N/A	Self-labeling	+Restaurant: 66.91%	+Easy to implement approach +Using different handcraft features	+Cannot utilize the correlation between categories +Imbalance problem during training +Depend on the feature selection approach +Two sentiment polarity classes
Thin et al. [41]	+Ensemble deep learning architectures +Joint training between aspect category with sentiment polarity Domain static word embedding	+Replace special words +Lowercase sentence +Word segmentation	VLSP2018	+Restaurant: 64.78% +Hotel: 79.90%	+Easy to implement approach +Utilize the correlation between categories +No handcraft feature enginree	+Require a large annotated dataset +Computational complexity +Not effective for data-less aspect categories
Tran and Bui [50]	+Hierarchical architecture based on BERT model +Fine-tuning the pre-trained language BERT model Multi-label classification	N/A	VLSP2018	+Restaurant: 71.30% +Hotel: 74.69%	+Utilize the correlation information of all layers +No handcraft feature enginree	+ Imbalance problem +Require a large annotated dataset +Computational complexity +No effectiveness for long input +Not effective for data-less aspect categories

* We denote the "self-labelling" value to indicate the datasets provided in the same studies.

Table 7. Summary of Machine Learning Approaches for the Aspect-Category Sentiment Analysis Task (Continued)

Papers	Method	Pre-processing steps	Datasets	Results	Advantages	Disadvantages
Lê et al. [22]	+Transfer learning based on BERT models. +Multilingual pre-trained language mBERT model +Multi-label classification	+Word normalization +Removing special characters, words +Sentence segmentation	VLSP2018	+Restaurant: 68.71% +Hotel: 67.72%	+Solving the document-level dataset based on the sentence-level training dat +Applying various processing steps +Utilize the power of BERT representation	+Imbalance problem during training +Require a large annotated dataset +Computational complexity +Not effective for data-less aspect categories +Require to annotate the sentence-level training data
Luc Phan et al. [21]	+Ensemble deep learning architecture CNN and Bi-LSTM models +Multi-label classification	N/A	Self-labeling	+SmartPhone: 63.06%	+Easy to implement approach +No feature engineer	+Imbalance problem during training +Computational complexity +Not effective for data-less aspect categories
Thin et al. [45]	+Transfer learning approach +Fine-tuning the BERT-based models. +Multi-task learning	N/A	Self-labeling	+Restaurant: 74.88% +Hotel:73.69%	+Employ the multi-task approach on BERT model +Utilize the correlation between categories	+Imbalance problem during training +Require a large annotated dataset +Computational complexity
Dang et al. [9]	+Transfer learning on BERT model +Multi-task learning	+Removing HTML tags +Standardize Unicode +Normalize Acronym +Word segmentation +Remove unnecessary character	VLSP2018	+Restaurant: 71.55% +Hotel: 77.32%	+Fine-tuning the contextual language models +Utilize the correlation between categories +No handcraft feature enginree	+Require a large annotated dataset +Computational complexity +No pre-processing for long input
Thin et al. [44]	+ Joint training architecture +Ensemble deep learning architecture	+Word normalization +Removing special characters, words +Word standardization +Correcting elongated words +Word segmentation	VLSP2018 KSE_2019	+Restaurant: 67.96% +Hotel: 73.16% +KSE2019: 69.13%	+Utilize the correlation information of all layers +No handcraft feature enginree	+ Imbalance problem +Require a large annotated dataset +Computational complexity +Not effective for data-less aspect categories
Le et al. [17]	+Transfer learning approach +Fine-tuning the BERT-based models. +Multi-task learning	+Lowercase +Meaningless characters removal Stopword removal	UITViSFD	+Smartphone: 78.76%	+Investigate various pre-trained BERT model	+Imbalance problem during training +Require a large annotated dataset +Computational complexity

* We denote the "self-labeling" value to indicate the datasets provided in the same studies.

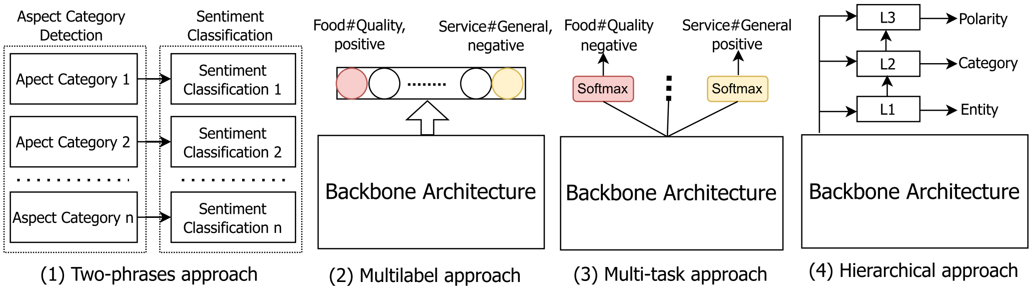

Fig. 4. Four types of architectures for the Vietnamese Aspect-Category Sentiment Analysis task.

The limitation of this approach is that the overall model performance depends on the Aspect Category Detection task. In addition, this straightforward approach ignores the relation information between categories with sentiments. Some early studies [9, 30, 41, 44, 45] explored the multi-task learning framework based on different classification algorithms to detect two tasks at the same time (see Figure 4(3)). Specifically, Reference [30] formatted the output of two tasks to a single one-hot vector and trained the SVM algorithms combined with handcraft features for each aspect category. Then, the authors trained multiple classifiers to predict the combination of category-sentiment pairs separately. However, training individual models will lose information about the relationship of occurrences between categories. To address this problem, References [38, 41] proposed an ensemble architecture based on BiLSTM and CNN models in the multitask manner by designing the number of output layers on the top of feature representation. The authors used the BiLSTM model to extract higher-level representations, then fed them into the CNN model with multiple output layers to identify the category-sentiment pairs. Besides, the authors also compared the effectiveness of different pre-trained word embedding methods for Vietnamese language.

Another work [50] also verified their proposed model on the document-level datasets. Their architecture is based on the BERT models and a hierarchical model to predict three-level labels (entities, attributes, and polarity) (see Figure 4(4)). For the BERT models, they exploited two versions, including **multilingual BERT (mBERT)** [10] and viBERT, which is further pre-trained mBERT on Vietnamese corpora. The classification model consists of three layers in the hierarchical structure that can capture the lower layers' information. Their experimental results indicated that monolingual BERT could improve the overall performance of multilingual models.

Instead of designing the architectures with multiple output layers, the authors of Reference [22] represented the number of category-sentiment pairs as a single vector and fine-tuned the mBERT model on two document-level VLSP2018 datasets (see Figure 4(2)). The output vector is formatted with the dimension of the number of categories multiplied by the number of sentiment polarity classes (e.g., for the restaurant with 12 aspect categories and 03 sentiment classes, the dimension of the output vector is 36 elements). This representation of the output leads to imbalanced training, especially for the domains with a large number of categories. Moreover, we found that the authors split the original datasets at the document level into sentences, then manually annotated the data in the training set. During the testing process, the authors also applied the sentence segmentation technique and combined the prediction on the single sentences to create the output for the document-level test set. The results showed that this approach is able to achieve better results in terms of F1 score for the restaurant domain, but the difference is not significant. Furthermore, training models on re-annotated training data and comparing them to previous studies does not provide a fair comparison.

Luc Phan et al. [21] also represented the output classes to a single multi-label vector and employed the ensemble deep learning based on BiLSTM and CNN architectures. To diversify the input features, the author concatenated the global average and max pooling on the top of the CNN model. The experimental results indicated that this approach could achieve good performances for the aspect categories with a high frequency of occurrence. Because this model is evaluated on their own manually annotated dataset and compared with traditional machine learning and deep learning models, it is difficult to conclude that this architecture is good for Vietnamese ABSA research. Another work [45] presented a comprehensive investigation to compare the performance of single learning and multitasking learning on two sentence-level datasets. Two learning strategies are evaluated based on different deep learning and transformer-based architectures. Their results demonstrated that the multi-task learning approach achieved better performance than the single-learning approach.

Similarly, instead of using deep learning models as the previous study [41], the work of Dang et al. [9] utilized the power of the pre-trained PhoBERT language model as a backbone model to extract the contextual representations and applied the multi-task strategy to train models. Moreover, their results showed that using the last four hidden state representations of the PhoBERT model gave a best performance than other representation approaches on two document-level datasets [29]. Nevertheless, the weakness of PhoBERT is the limitation of the length of the input, where this model only supports the input with a maximum length of 256 tokens. This results in a huge loss of information for long comments, especially document-level sentences, because the model applies the truncation method automatically.

Le et al. [17] presented a study to investigate the performance of fine-tuning the pre-trained language models on the UITViSFD dataset for the smartphone domain. The proposed model is designed by adding the output layer on top of the last layer of the pre-trained language PhoBERT models and then fine-tuning directly on the training set. To represent the list of aspect categories and their corresponding sentiment analysis, the output layer is a multi-label vector, with the length of the vector as the number of categories multiplied by the number of polarity classes. The authors employed different transformer-based pre-trained language models as the backbone, including BERT, XLM-R, RoBERTa, and DistilBERT. The experimental results showed that fine-tuning the PhoBERT model outperformed other approaches. Their model achieved the 78.76% in terms of macro F1-score, which is higher than the baseline result [21], approximately 24.90%.

Thin et al. [44] proposed a joint architecture that utilized the additional information from the ACD task to improve the performance of the ACSA task. Their results showed that the joint training of these two model paradigms enhanced the performance and significantly allowed them to outperform existing SOTA scores on three document-level benchmark datasets. Specifically, the authors employed two deep learning architectures (BiLSTM-CNN and BiGRU). This article argued that the BERT model limits the length of the input; therefore, the authors used static domain word embedding combined with deep learning the learn the feature representation for the document-level input. Their experimental results demonstrated that the proposed architecture outperformed previous methods and baseline models on three document-level benchmark datasets.

Recently, based on the power of contextual pre-trained language models, which are available for Vietnamese language, the work of Thin et al. [43] presented an ensemble framework to solve the ACSA task on three benchmark datasets. Their architecture used the soft-voting technique to combine the output of based classifiers. To train the based classifiers, the authors combined the multi-task approach to BERTology models, including the multilingual model and monolingual models. The experiments were conducted on the UIT_ABSA and UIT-ViFSD datasets, respectively. Their results showed that the ensemble model could improve the overall performances compared with the based classifiers and previous studies. In our perspective, this approach completely depends on the BERTology models as a backbone, and the computational cost is relatively high to combine the output of different based classifiers.

4.2.3 Aspect-Term Sentiment Analysis. For the Aspect-Term Sentiment Analysis task, the work of Mai and Le [23] is the first attempt, which has been to study this task in Vietnamese language. The authors proposed a sequence-to-sequence scheme based on different bidirectional recurrent neural networks combined with CRF models. To address the sequence labelling problem, each token in a sentence will be assigned a specific tag to represent the aspect target and its sentiment polarity classes through BIO format. Three variants of RNN models are used to extract the sentence representation, including the simple RNN, LSTM, and GRU models. Moreover, the authors investigated the effectiveness of pre-trained word embeddings in deep learning architectures. Their experimental results BGRU-CRF model on the domain-specific word embedding achieved the highest F1-score in terms of 71.79%. Following that, Reference [24] leveraged the knowledge from the sentence sentiment analysis task to improve the performance on the ATSA task by joint training two tasks in the same architecture. The authors assumed there was a correlation between the sentiment of all aspects and the overall sentiment of the sentence. In their paper, many sequence labeling architectures were employed on their annotated dataset, including three deep learning architectures of Reference [23], CRF model, and pre-trained BERT model. The result showed that fine-tuning the pre-trained language model BERT gave the best performance on the test set. However, the author only employed the multilingual BERT instead of using other monolingual models, which can produce a better contextual representation of Vietnamese words.

4.2.4 Aspect Sentiment Classification. To predict the sentiment class for each aspect category mentioned in the sentence, Nguyen et al. [31] applied the Convolutional Neural Network architecture with multiple filter sizes combined with pre-trained word2vec as a baseline for the sentiment classification task. Because their paper focuses on conducting data preprocessing at the sentence-level dataset, they just reported the results on different types of datasets, such as standardized or not standardized, in different pre-processing layers. It is difficult to conclude that this approach is suitable for the ASC task, because the authors conduct the experiments on their creating datasets. With the same idea as the previous study [31], instead of using the CNN architecture as the base classifier, the work of Ngoc et al. [27] explored the performance of the ensemble CNN-LSTM model. The authors also experimented on the same dataset, which was presented in Reference [31]. Their experimental results indicated that the CNN+LSTM model was able to achieve promising results with a small corpus.

Le-Minh et al. [19] proposed a Mini-window Locating Attention technique to select the aspect category and sentiment words in the review. This method is expected to ignore irrelevant words from target classes based on the importance weight of words. To calculate these values, the Chi-Squared technique is used to estimate the score. For the classification component, the authors employed various traditional machine learning algorithms, including SVM, Decision Tree, Naive Bayes, Random Forrest, and Logistics Regression, as the classifiers and trained each classifier for each category-sentiment pair separately. Compared to proposed representations with other approaches, such as BERT and one-hot encoding, their method outperformed two self-annotated datasets. However, this approach was highly dependent on choosing a manual threshold to calculate the word's score. Moreover, the authors did not compare the performance with the transformer-based architectures to demonstrate the effectiveness of the proposed method.

4.2.5 Span Detection. Span Detection task is one of the challenging tasks in the ABSA field. The purpose of this task is to extract one or more spans of aspect categories and their sentiments in the review. To the best of our knowledge, Thanh et al. [40] is the first attempt to tackle this task on Vietnamese review. The authors considered the Span Detection task as a sequence labeling problem and proposed a BiLSTM-CRF architecture with fusion embedding. The fusion embedding layer consists of three representations from three pre-trained models, including syllable embedding, character embedding, and contextual embedding from XLM-R. After obtaining the final representation of the input in the BiLSTM layer, a CRF model is added on top of the BiLSTM model to calculate the conditional log-likelihood of tag sequences. To convert the output for the sequence, the **BIO format (Beginning, Inside, and Outside)** was used to represent the aspect category task and corresponding sentiment polarity. Although this architecture provided promising results, there were some points that can help future research on this dataset. First, the model often gave the wrong predictions on the spans that contain the English loanwords. Second, the data was annotated in the spans of the strings; therefore, it can contain some noisy characters such as punctuation, stopwords, or icon symbol. Moreover, the model fails to predict the "Neutral" classes for the aspect category.

Recently, Le et al. [16] proposed a hierarchical model combined with multi-task learning to address the span detection task and aspect category sentiment analysis at the same time. To train joint architecture, the author presented a controllable task-dependent loss for their hierarchical multi-task model. Their work used the IOB scheme to represent the span sequence, aspect category, and sentiment polarity class as two formats, including span with category and span with polarity. Therefore, the proposed model consists of two blocks according to two scheme representations. The pre-trained language PhoBERT model was employed to produce the contextualized representation for the input. Then a CRF algorithm was added to predict the label sequence of the aspect category. Then, the feature representation of the category block was concatenated with the feature of the polarity block to learn the hierarchical information between tasks. The controllable loss was proposed to optimize the model. The experiments were conducted on the UIT-ViSD4SA and demonstrated the effectiveness compared with the baseline system [40]. Nevertheless, this approach cannot solve the imbalance problem in the training data, and the configuration of controlled task-dependent loss was chosen manually.

4.2.6 Aspect Term Extraction. To tackle this task, Reference [18] introduced a semi-supervised approach based on the topic modeling to aspect extraction from product comments. The algorithm GK-LDA is employed to build a model with prior knowledge from lexical relation sets, which contain the relationship of aspect categories and terms. After extracting the terms, three measures, including the aspect frequency, cosine similarity, and decision tree, were used to classify the aspect categories. This approach is evaluated on a manual test set and reported through the Accuracy metric. The experimental results indicated that this approach is ineffective for samples containing slang words and precise meanings about a specific aspect. Moreover, this approach requires the linguistic knowledge of experts to create the sets of initial aspect terms and the manual processing steps to address the slang words and words with explicit meanings.

4.3 Datasets

After analyzing the 31 selected manuscripts, we found that many benchmark datasets were annotated for the same tasks and level of data except for the domain. Table 11 presents the information on available published datasets for Vietnamese ABSA research. For example, the VLSP2018 [29], UIT-ViFSD [21], and PACLIC2022 [51] were the document-level datasets for the ACD and ACSA tasks for the different domains, such as restaurant, hotel, smartphone, and beauty, respectively. Moreover, their works used the approaches that have been proposed by other studies as the baseline to evaluate the new datasets; therefore, it can limit the development of this topic in Vietnamese. Although annotating a new dataset for a specific domain is necessary because it might be useful in real-world applications. However, this led to difficulty in comparing different methods and architectures on the benchmark datasets in terms of scientific research. Therefore, this section summarizes the benchmark datasets that were annotated and released for the NLP community for Vietnamese language. Furthermore, we provide statistical information as well as current SOTA

Table 8. The Results of Previous Studies on Two VLSP2018 Datasets

Approach	Restaurant		Hotel	
	ACD	ACSA	ACD	ACSA
SVMs [46]	77.00	61.00	70.00	61.00
DCNN [42]	80.40	-	69.25	-
BiLSTM-CNN [41]	79.70	64.78	77.85	70.90
Hier-BERT [50]	84.23	71.30	82.06	74.69
BERT [22]	81.35	64.14	79.66	67.74
Joint BiLSTM-CNN [44]	82.29	67.96	78.66	73.16
Multitask-BERT [9]	83.29	71.55	82.55	77.32
HSUM-HC [52]	-	-	85.20	80.08

*We reported the best F1-scores in previous studies.

performance on publicly available benchmarks. This information helps the reader have an overview of public datasets and would push forward the research in the Vietnamese ABSA field.

To the best of our knowledge, no public dataset has been released for the research community before 2018. At the shared-task VLSP 2018, Reference [29] presented two document-level datasets for the ACD and ACSA tasks for two domains. There are 12 and 34 types of aspect categories for the restaurant and hotel domains, respectively. Each aspect category is assigned one of three sentiment classes: positive, negative, and neutral. We observed that they are challenging datasets for the following reasons: (1) there is an imbalance problem between aspect categories and sentiment polarities; (2) the average length of samples in the test set is longer than the training set; (3) containing a lot of typographical errors, syntax, noise elements in the data. Table 8 reported the F1-scores of previous studies on two benchmarks.

Another benchmark dataset was provided in Nguyen et al. [30] for the ACD and ACSA tasks. The authors crawled the data from the Foody site, which is a famous website about restaurants in Vietnamese. In total, this dataset consists of 7,828 reviews, where each review can be assigned seven types of aspect categories and five levels of sentiment polarities. This dataset is labeled by three annotators with high inter-annotator agreements. The authors also presented a baseline approach based on multiple SVM classifiers with handcraft features. In general, this baseline achieved the F1-score of 87.13% and the F1-score of 59.20% for the ACD and ACSA tasks, respectively.

Instead of labeling the dataset in the document-level review, Reference [45] built and released two sentence-level datasets for the hotel and restaurant domain. It was also annotated following the guidelines and scheme of VLSP competitions 2018 [29]. As a result, nearly 10,000 samples for each domain are released for the community with the free-purpose research. The authors also applied the iterative stratification technique to split datasets into three sets. This technique helps the dataset balance the class ratio during data division. The datasets are divided into three sets with a ratio of 7/1/2. For the restaurant domain, the best F1 score was 86.96% and 74.88% for ACD and ACSA, respectively. The best model achieved an F1-score of 79.10% and 73.69% for the hotel domain. Table 9 presented the SoTA results of previous studies on two datasets.

Luc Phan et al. [21] also presented a new benchmark Vietnamese dataset (named UIT-ViFSD) for the mobile phone e-commerce domain, consisting of 11,122 human-annotated comments with the ACD with ACSA tasks. This dataset is also published freely for research purposes. However, compared to datasets in other languages for the same domain, this dataset has 10 aspect categories instead of 17 aspect categories as previous datasets [37]. Furthermore, the limitation of having different numbers of labels and class names compared to international datasets will limit some research directions on multilingual problems such as cross-lingual ABSA, zero-shot learning, and so

Table 9. The Results of Previous Studies on Two UIT_ABSA Datasets

Approach	**Restaurant**		**Hotel**	
	ACD	**ACSA**	**ACD**	**ACSA**
BERT [53]	86.53	-	79.16	-
Mulitask BERT [45]	86.96	74.88	77.66	73.69
HSUM-HC [52]	-	-	80.78	75.25
Ensemble BERTology [43]	-	75.28	-	73.77

*We reported the best F1-scores in previous studies.

Table 10. The Results of Previous Studies on Two UIT-ViFSD Datasets

The work of	**Approach**	**Model**	**Task**	
			ACD	**ACSA**
[21]	Traditional machine learning	Naive Bayes	64.65	37.56
		SVM	42.63	19.69
		Random Forrest	47.83	20.17
	Deep learning	CNN	69.70	27.16
		LSTM	80.27	52.13
		BiLSTM-CNN	84.48	63.06
[17]	Transformer-based learning	XLM-R	82.73	65.81
		DistilBERT	84.56	62.67
		RoBERTa	66.26	50.88
		BERT	72.56	40.64
		PhoBERT	86.03	78.76
[43]	Ensemble learning	Multilingual models	-	75.83
		Monolingual models	-	74.93
		Two best models	-	76.07
		Five models	-	76.27

on. The BiLSTM-CNN model combined with the output's representation as a multi-label problem achieved the baseline results with an F1-score of 84.48% and an F1-score of 63.03% for the ACD and ACSA tasks, respectively. Different from previous studies [29, 45], this dataset is measured on the macro-average scores instead of the micro-average scores. Table 10 shows the performance of different approaches in this dataset.

Another dataset for the education domain is presented by Sau et al. [38]. Similar to previous works, this dataset is labeled for two tasks in ABSA research, including the ACD and ACSA on the student's feedback of a university. The dataset consists of 5,010 sentences, where each sample is assigned from 11 different aspect categories and 03 sentiment classes. Because the dataset has 03 sentiment classes, including "positive", "negative", and "neutral" label. Based on the data distribution statistics, we observed a high imbalance between sentiment classes for aspect categories, especially the "neutral" class. The best model from experiments is the BiLSTM-CNN model with a micro F1-score of 78,93% and 73.78% for the ACD and ACSA tasks, respectively. The dataset was also provided for the research community via contacting the authors to access the data.

Le-Minh et al. [19] also presented a dataset that is crawled from two e-commerce platforms for the ACD and ACSA tasks. The reviews are assigned to 8 and 6 different aspect categories for the technology and Mother&Baby domain, respectively. Different from previous works [29, 30, 45], the author just used the two sentiment classes (positive and negative) instead of three or five labels. Different models were also implemented on two datasets to provide the baseline results for

Table 11. Summary of Vietnamese Benchmark Datasets for ABSA Tasks

Papers	Dataset*	Level	Domain	Source	Coverage Task	Dataset Size	Availability of Dataset	Annotation Agreement	Test Phase	Evaluation Metrics
Nguyen et al. [29]	VLSP2018	Document	Restaurant	Foody	ACD ACSA	4,751	Yes[a]	N/A	Yes	Micro F1-score
Nguyen et al. [29]	VLSP2018	Document	Hotel	Agoda	ACD ACSA	5,600	Yes[b]	N/A	Yes	Micro F1-score
Thuy et al. [47]	KSEO2018	Sentence	Restaurant	Foody	ACD	-	No	83.00%	CV	Micro F1-score
Mai and Le [23]	ACIIDS2018	Sentence	Smartphone	Youtube	ATSA	2,098	No	N/A	Yes	F1-score
Nguyen et al. [30]	UIT_ABSAv1	Document	Restaurant	Foody	ACD ACSA	7028	Yes[c]	ACD: 92.87% ACSA: 88.75%	Yes	Micro F1-score
Nguyen et al. [31]	ACIIDS2019	Sentence	Car Product	N/A	ASC	4,829	No	N/A	Yes	Weighted F1-score
Thuy et al. [48]	RIVF2020	Sentence	Restaurant	Foody	ACSA	-	No	86.00%	CV	Micro F1-score
Thin et al. [45]	UIT_ABSAv2	Sentence	Restaurant	Foody	ACD ACSA	9,737	Yes[d]	ACD: 86.32% ACSA: 77.63%	Yes	Micro F1-score
Thin et al. [45]	UIT_ABSAv2	Sentence	Hotel	Mytour	ACD ACSA	10,002	Yes[e]	ACD: 85.53% ACSA: 82.01%	Yes	Micro F1-score
Luc Phan et al. [21]	UIT-ViSFD	Document	Smartphone	E-commerce	ACD ACSA	11,122	Yes[f]	ACD: 73% ACSA: 72%	Yes	Macro F1-score
Le-Minh et al. [19]	VECD	Document	Technology	Tiki Shopee	ACD ACSA	6,238	Yes[g]	N/A	CV	Macro F1-score Micro F1-score
Le-Minh et al. [19]	VECD	Document	Mother&Baby	Tiki Shopee	ACD ACSA	6,002	Yes[h]	N/A	CV	Macro F1-score Micro F1-score
Sau et al. [38]	TNU2021	Sentence	Education	University	ACD ACSA	5,010	Yes[i]	88.95%	Yes	Micro F1-score
Thanh et al. [40]	UIT-ViSD4SA	Document	Smartphone	E-commerce	SP	11,122	Yes[j]	N/A	Yes	Micro F1-score Macro F1-score
Mai and Le [24]	AORJ2021	Sentence	Smartphone	Youtube	ATSA	2,153	Yes[k]	N/A	Yes	F1-score
Tran et al. [51]	PACLIC2022	Document	Beauty	Shopee	ACD ACSA	16,227	Yes[l]	N/A	Yes	Weighted F1-score

In case the dataset is not named, we use the conference name with the year to name the dataset. CV is the k-fold cross-validation technique.

[a] https://vlsp.org.vn/resources-vlsp2018
[b] https://vlsp.org.vn/resources-vlsp2018
[c] https://sites.google.com/uit.edu.vn/uit-nlp/datasets-projects?authuser=0
[d] https://github.com/undertheseanlp/underthesea
[e] https://github.com/undertheseanlp/underthesea
[f] https://github.com/LuongPhan/UIT-ViSFD
[g] https://github.com/phuonglt26/Vietnamese-E-commerce-Dataset
[h] https://github.com/phuonglt26/Vietnamese-E-commerce-Dataset
[i] Contact to the first author.
[j] https://github.com/kimkim00/UIT-ViSD4SA
[k] https://github.com/mailong25/sentiment-analysis
[l] Contact to the first author.

future studies. However, the authors just conducted the experiments on the ASC only, no baseline approach was presented for the ACD or ACSA tasks.

The work of Tran et al. [51] presented a new dataset at the document-level review for the beauty domain. This dataset is also annotated for the ACD and ACSA tasks simultaneously. There are seven types of aspect categories and one spam class. Each category is assigned to one of three sentiment polarity classes: "positive," "negative," and "neutral." In total, this dataset consists of 16,227 reviews and is divided into three sets with a ratio of 8/1/1 for training, development, and testing sets. Unfortunately, we found that there is a serious imbalance problem between sentiment classes of all aspect categories. For example, the category "prices" has 3,238 samples for "positive," but the "negative" and "neutral" classes just have only 21 and 27 samples, respectively. The BiGRU model combined with the Convolution layer achieved the best performance on the testing set with 97.51% of weighted F1-score for the ACD task and 86.92% of weighted F1-score for the ACSA task. The highest performance is gained by the "positive" class due to the imbalance problem in the dataset. Therefore, future work should consider the imbalanced sentiment classes to enhance the overall performance of this dataset.

For the Aspect-Term Sentiment Analysis task, Mai and Le [24] published a benchmark dataset on comments from commercials and videos regarding smartphones. Each sentence is manually assigned to possible terms-sentiment pairs in the sentence. Moreover, the annotators are required to label the overall sentiment of the review. The aspect term is defined as a word or multi-words that indicate an entity (e.g., product names) or features of smartphones (e.g., camera, battery). Totally, this dataset consists of 2,153 sentences with 3,103 aspect terms-sentiment pairs. The author conducted massive experiments on this dataset and showed that joint training models between ATSA and sentence sentiment tasks based on fine-tuning the BERT model achieved the best scores for two tasks. Finally, the best F1-scores for the ATSA task is 81.78% on the test set.

For the Span Detection task, Reference [40] released a new Vietnamese dataset (UIT-ViSD4SA) by annotating the target spans for each aspect category of the UIT-ViFSD dataset [21]. Unlike the Opinion Target Extraction task [37], each span is detected as a continuous sequence of words in the review, which express the explicit or implicit aspect categories and their sentiments. Therefore, the lengths of spans in the dataset are different. In total, 35,396 spans are labeled for 11,122 real-world smartphone reviews. The author treated this task as a sequence labeling problem for the baseline model and employed the BiLSTM-CRF combined with fusion embedding. The results are reported in the micro and macro scores for two subtasks: span for the aspect category, span for the sentiment polarity, and span for the category-sentiment pair. Precisely, the best performance of detecting spans for category-sentiment pairs is 62.38% in the micro F1-score and 45.70% in the macro F1-score.

Tables 12 and 13 show the statistics about publicly available datasets for the ACD and ACSA tasks at the sentence-level and document-level review, respectively. The information includes the number of samples, the number of labels, the average length and maximum length, the number of vocabulary and the number of aspect categories, and the ratio of average aspect categories in the whole dataset. This information can provide an overview of benchmark datasets for future work on the Vietnamese ABSA field.

4.4 Evaluation Metrics

This section summarizes the evaluation metrics for each task in the Vietnamese ABSA research.

For the span detection, previous studies [16, 40] reported the Precision, Recall, and F1-score on both the micro and macro averages. A predicted span with the aspect category and polarity is calculated only if it exactly matches the corresponding gold span. For the Aspect Category Detection task and Aspect-Category Sentiment Analysis task, most previous studies [29, 30, 41] used

Table 12. Statistics about the Publicly Available Sentence-level Datasets for the ACD and ACSA Task

Information	Datasets		
	Restaurant	Hotel	Education
Number of samples	9,737	10,005	5,010
Number of labels	12 × 3	34 × 3	11 × 3
Average Length	16.9	18.3	12.76
Maximum Length	96	192	147
Vocabulary size	16,773	8,041	2,378
Number of aspect categories	13,140	16,413	6,076
Average Aspects/Samples	1.40	1.64	1.21

Restaurant and Hotel: [45].
Education: [38].

Table 13. Statistics about the Publicly Available Document-level Datasets for the ACD and ACSA Task

Information	Datasets						
	Restaurant	Phone	Hotel	Beauty	Technology	Mo&Ba	Restaurant*
Number of samples	4,751	11,122	5,600	16,227	6,238	6,002	7,828
Number of labels	12 × 3	10 × 3	34 × 3	8 × 3	8 × 2	6 × 2	7 × 5
Average Length	66.09	36.2	45.82	20.89	12.76	14.72	52.07
Maximum Length	874	250	1,022	122	147	199	208
Vocabulary size	9,783	8,063	6,886	13,500	2,378	1,974	7,521
Number of aspect categories	14,861	33,710	23,643	32,791	9,586	7,046	24,653
Average Aspects/Samples	3.19	3.3	4.22	2.01	1.92	1.62	3.51

Restaurant and Hotel: [29], Phone: [21], Beauty: [51].
Technology and Mother&Baby: [19], Restaurant*: [30].

micro-averaged precision, recall, and F1-score to evaluate the models. Three metrics are calculated as follows:

$$Precision = \frac{\sum_{c_i \in C} TP_{c_i}}{\sum_{c_i \in C} TP_{c_i} + FP_{c_i}}, \tag{1}$$

$$Recall = \frac{\sum_{c_i \in C} TP_{c_i}}{\sum_{c_i \in C} TP_{c_i} + FN_{c_i}}, \tag{2}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \tag{3}$$

where TP_{c_i} is the number of correct predictions for the category c_i, FP_{c_i} is the number of incorrect predictions for the category c_i, FN_{c_i} is the number of ground truth predictions, which is not assigned for the category c_i, and c_i is the ith category in the number of C aspect categories. The score for the ACSA task is calculated when a category-sentiment pair is correctly predicted. However, we found that a few studies used other measures to evaluate the ACD and ACSA tasks. For example, Luc Phan et al. [21] reported the results on the macro-average scores, while the work of Tran et al. [51] used the weighted scores to evaluate the results.

5 DISCUSSION AND RESEARCH DIRECTIONS

This section presents a discussion of the challenge from previous studies and suggests some possible future research directions for this topic in Vietnamese language.

Imbalance problem: When analyzing the statistics of several benchmark datasets, it is obvious to consider the imbalance problem between aspect categories as well as sentiment polarity classes. We found that a few existing studies solved this problem; however, their approaches are quite simple and cannot be applied to general datasets. Specifically, Reference [47] applied the back-translation technique to expand the training data. Although this approach can improve overall performance, the difference is not significant. Therefore, future works can focus on the data augmentation techniques [7, 11] based on the characteristics and style of comments in Vietnamese to handle the imbalance problem.

Informal style of writing and grammatical errors: One of the biggest challenges is the style of writing and grammatical errors in the users' reviews. We found that people tend to use acronyms (e.g., (food) "đồ ăn," "thức ăn," "món ăn," or "món nhậu"), emoticons (e.g., >.<, .), abbreviations (e.g., the word "không" (no) can be written as "k," "hog," "ko," etc.), dialects, and non-standard syntax to express their opinions in the review. Therefore, word segmentation tools perform poorly on such user reviews, because these tools are trained on a standard text corpus. Moreover, the rate of comments with misspellings or words without tone marks appeared in many datasets. For example, given a review on YouTube as "điện thoai mơi lạ nhìn dep" (correct: điện thoại mới lạ nhìn đẹp—"The new phone looks unique and beautiful"). It is very difficult to build a general model to solve the spelling mistake of users. With the non-accent words in the review, future studies can employ the accent restoration model [35] with the power of sequence-to-sequence architectures to solve this problem for a specific domain.

Code mixing review: Recently, the code-mixing problem has appeared with increasing frequency in user comments due to the development of multilingual environments. Code-mixing or code-switching is the phenomenon of mixing the vocabulary and syntax of two or more languages in the same sentence [15]. As we know, the development of pre-trained language models has brought a new revolution in NLP tasks, including Aspect-based sentiment analysis. However, applying the pre-trained language models, including the multilingual or monolingual model on the code-mixed review, degrades the overall performance [39]. For example, the comment "giá thức ăn thì high" (the price of the food is high) is written in a code-mixed style. Some previous studies designed a dictionary to correct the code-mixed vocabulary to Vietnamese meanings in the pre-processing step to tackle this challenge. However, this approach is limited to a specific dataset and cannot be applied in real-world applications.

Lack of availability of comprehensive benchmark datasets: The availability of benchmark datasets is crucial for the development of the Vietnamese ABSA research field. However, we indicated that all published datasets were only annotated simultaneously for one or two tasks. Moreover, we found that there are some current studies focusing on building existing tasks on other data domains instead of annotating the new tasks on the available published dataset. For example, the previous works [21, 51] annotated the same ACD and ACSA tasks for the beauty, phone domains, respectively. This will limit the development of methods, models, and research directions of the ABSA field in Vietnamese. Recently, the latest studies [2, 55] in other languages released benchmarks that are annotated many tasks in ABSA research. The comprehensive benchmark datasets are essential, because they would push the research in this field forward. Therefore, we recommend that future research should annotate the other tasks on the published datasets instead of labelling the existing tasks on other domains.

Limitation of existing approaches, methods, and algorithms: From our analysis in Section 4, most existing methods are based on machine learning, deep learning, and transformer-based learning combined with two-stages approach or multi-task approach to address the tasks in Vietnamese ABSA benchmarks. However, we found that a few studies presented new approaches to the public dataset. Instead, they published datasets on the new domains and presented the baseline results from the approaches of previous studies. However, these baseline approaches are quite outdated compared with the development of ABSA research in the world. We recommend some new approaches that can be studied for Vietnamese datasets. For example, converting the ABSA as a machine reading comprehension problem [25, 54], transforming the ABSA tasks as the text generation problem [6, 55], and utilizing the power of new pre-trained language models for Vietnamese such as ViT5 [36], PhoBART [49]. Furthermore, utilizing the in-context learning approach based on the latest large language models, such as ChatGPT or LLaMa, across different datasets can be a promising research direction.

Limitation of consistency in evaluation metrics: Based on the information in Sections 4.4 and 4.3, we can see that the previous studies reported the performance on the same tasks using different evaluation metrics. For example, References [29, 45] used the micro-average F1-score for the overall results like studies in other languages [37]. While References [19, 21, 51] used the macro-average and weighted F1-score to measure the same tasks. This will lead to ambiguity and inconsistency in the Vietnamese research community, especially for young researchers. Therefore, we suggest that future work verify the experimental datasets' evaluation metrics to compare methods fairly.

Limitations of utilizing the Vietnamese linguistic characteristics: Vietnamese has a diverse lexical system such as compound words, onomatopoeia, and pictograms, and the richness in the system of pronouns and phrases; therefore, the ways of expressing sentiment are also very diverse. Moreover, there are many different forms of comments in direct and indirect manners. For example, in an indirect way, we can use complimentary metaphors, such as compliments based on a question form (Sao nhà hàng có thể nấu được món ngon như thế này?—"How could the restaurant make such delicious dishes?") to express the sentiment to an aspect. It can be seen that the study of the ABSA problem in Vietnamese needs to pay attention to the language characteristics and commenting culture. Therefore, future research needs to focus on the comments' linguistic features to improve the performance of classifying the sentiment polarity for each aspect. Furthermore, it is vital to analyze the praise and criticism characteristics in Vietnamese culture to build standard datasets for ABSA problems in real-practice scenarios.

6 CONCLUSION

This article adopted a systematic literature review approach to review aspect-based sentiment analysis research in the Vietnamese language of 31 manuscripts. We provided an overview of the recent progress of this field in Vietnamese, including the task coverage, approaches, well-known benchmark datasets, and the gap and limitations of current research.

The 31 manuscripts are first categorized and summarize the research progress to specific tasks listed in the approach section (ACD, ACSA, ATSA, SC, SD, and ATE). The advantages and disadvantages of different existing approaches are analyzed and discussed in detail in the corresponding sections. Then, the information and statistics of publicly available benchmark datasets are provided for future studies. Moreover, the latest SOTA scores on the benchmark datasets are collected. Finally, we discuss the research challenge and provide the future directions for this field in Vietnamese language. Our article can act as a foundation to update researchers about the recent progress of the field in Vietnamese.

APPENDIX

A ACRONYMS

This section provides all acronyms used in the whole article.

Acronym	Meaning	Acronym	Meaning
ABSA	Aspect-based sentiment analysis	ACSD	Aspect Category Sentiment Detection
ATE	Aspect Term Extraction	ASQP	Aspect Sentiment Quad Prediction
ACD	Aspect Category Detection	TF-IDF	Term Frequency–Inverse Document Frequency
OTE	Opinion Term Extract	VLSP	Vietnamese Language and Speech Processing
ASC	Aspect Sentiment Classification	SMOTE	Synthetic Minority Oversampling TEchnique
AOPE	Aspect-Opinion Pair Extraction	SVM	Support Vector Machine
ATSA	Aspect-Term Sentiment Analysis	CNN	Convolution Neural Network
ACSA	Aspect-Category Sentiment Analysis	LSTM	Long short-term Memory
ASTE	Aspect Sentiment Triplet Extraction	BERT	Bidirectional Encoder Representations from Transformers
SOTA	State of the art	MLP	Multi-layer Perceptron
GRU	Gated Recurrent Unit	BiLSTM	Bidirectional Long short-term Memory
NLP	Natural Language Processing	CRF	Conditional Random Field
BIO	Begin-Inside-Outside	RNN	Recurrent Neural Network

ACKNOWLEDGMENTS

We would like to thank the reviewers and the chairs for their insightful reviews and suggestions.

REFERENCES

[1] Mark Alves. 2006. Linguistic research on the origins of the Vietnamese language: An overview. *J. Viet. Stud.* 1, 1-2 (2006), 104–130.

[2] Xiaoyi Bao, Wang Zhongqing, Xiaotong Jiang, Rong Xiao, and Shoushan Li. 2022. Aspect-based sentiment analysis with opinion tree generation. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI-')*, Lud De Raedt (Ed.). International Joint Conferences on Artificial Intelligence Organization, 4044–4050. DOI: https://doi.org/10.24963/ijcai.2022/561 Main Track

[3] Rajae Bensoltane and Taher Zaki. 2022. Aspect-based sentiment analysis: An overview in the use of Arabic language. *Artif. Intell. Rev.* 56 (2022), 2325–2363. DOI: https://doi.org/10.1007/s10462-022-10215-3

[4] Gianni Brauwers and Flavius Frasincar. 2022. A survey on aspect-based sentiment classification. *ACM Comput. Surv.* 55, 4, Article 65 (Nov. 2022), 37 pages. DOI: https://doi.org/10.1145/3503044

[5] Thanh Hung Bui and Quoc Binh Nguyen. 2020. Aspect-based sentiment analysis using hybrid deep learning approach CNN-LSTM. In *Proceedings of the 13th National Conference on Fundamental & Applied Information Technology Research*. Publishing House of Natural Science and Technology, 456–460. DOI: https://doi.org/10.15625/vap.2020.00200

[6] Siva Uday Sampreeth Chebolu, Franck Dernoncourt, Nedim Lipka, and Thamar Solorio. 2021. Exploring conditional text generation for aspect-based sentiment analysis. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*. Association for Computational Lingustics, 119–129. Retrieved from https://aclanthology.org/2021.paclic-1.13

[7] David Z. Chen, Adam Faulkner, and Sahil Badyal. 2022. Unsupervised data augmentation for aspect based sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, 6746–6751. Retrieved from https://aclanthology.org/2022.coling-1.586

[8] Minh-Son Cong, Tuan-Minh Dam, Anh-My Phuong, Thi-Quyen Nguyen, Duy-Cat Can, Hoang-Quynh Le, Long Hoang, and Mai-Vu Tran. 2021. THANOS: The aspect classification model for imbalanced Vietnamese E-commerce review data. In *Proceedings of the International Conference on Asian Language Processing (IALP'21)*. IEEE, 352–357. DOI: https://doi.org/10.1109/IALP54817.2021.9675163

[9] Hoang-Quan Dang, Duc-Duy-Anh Nguyen, and Trong-Hop Do. 2022. Multi-task solution for aspect category sentiment analysis on Vietnamese datasets. In *Proceedings of the IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom'22)*. IEEE, 404–409. DOI: https://doi.org/10.1109/CyberneticsCom55287.2022.9865479

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 4171–4186. DOI: https://doi.org/10.18653/v1/N19-1423

[11] Ting-Wei Hsu, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. Semantics-preserved data augmentation for aspect-based sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 4417–4422. DOI: https://doi.org/10.18653/v1/2021.emnlp-main.362

[12] Bui Thanh Hung, Vijay Bhaskar Semwal, Neha Gaud, and Vishwanth Bijalwan. 2021. Hybrid deep learning approach for aspect detection on reviews. In *Proceedings of Integrated Intelligence Enable Networks and Computing*, Krishan Kant Singh Mer, Vijay Bhaskar Semwal, Vishwanath Bijalwan, and Rubén González Crespo (Eds.). Springer, 991–999.

[13] Quang Thắng Đinh, Hồng Phương Lê, Thị Minh Huyền Nguyễn, Cẩm Tú Nguyễn, Mathias Rossignol, and Xuân Lương Vũ. 2008. Word segmentation of Vietnamese texts: A comparison of approaches. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2008/pdf/493_paper.pdf

[14] Barbara Kitchenham. 2004. Procedures for performing systematic reviews. *Keele, UK, Keele University* 33, 2004 (2004), 1–26. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=29890a936639862f45cb9a987dd599dce9759bf5

[15] Yash Kumar Lal, Vaibhav Kumar, Mrinal Dhar, Manish Shrivastava, and Philipp Koehn. 2019. De-mixing sentiment from code-mixed text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics, 371–377. DOI: https://doi.org/10.18653/v1/P19-2052

[16] An Pha Le, Tran Vu Pham, Thanh-Van Le, Duy V. Huynh, and Dat Ngo. 2022. Robust hierarchical model for joint span detection and aspect-based sentiment analysis in Vietnamese. In *Proceedings of the 9th NAFOSTED Conference on Information and Computer Science (NICS'22)*. IEEE, 35–40. DOI: https://doi.org/10.1109/NICS56915.2022.10013463

[17] Bao Hoang Le, Hoang Minh Nguyen, Nhi Kieu-Phuong Nguyen, and Binh T. Nguyen. 2022. A new approach for Vietnamese aspect-based sentiment analysis. In *Proceedings of the 14th International Conference on Knowledge and Systems Engineering (KSE'22)*. IEEE, 1–6. DOI: https://doi.org/10.1109/KSE56063.2022.9953759

[18] Hai Son Le, Thanh Van Le, and Tran Vu 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

[19] Binh Le-Minh, Thi-Phuong Le, Khanh-Hung Tran, Khanh-Huyen Bui, Hoang-Quynh Le, Duy-Cat Can, Hung Nguyen Chung Thanh, and Mai-Vu Tran. 2021. Aspect-based sentiment analysis using mini-window locating attention for Vietnamese E-commerce reviews. In *Proceedings of the 13th International Conference on Knowledge and Systems Engineering (KSE'21)*. IEEE, 1–4.

[20] Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthes. Lect. Hum. Lang. Technol.* 5, 1 (2012), 1–167.

[21] Luong Luc Phan, Phuc Huynh Pham, Kim Thi-Thanh Nguyen, Sieu Khai Huynh, Tham Thi Nguyen, 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 *Knowledge Science, Engineering and Management*, Han Qiu, Cheng Zhang, Zongming Fei, Meikang Qiu, and Sun-Yuan Kung (Eds.). Springer International Publishing, Cham, 647–658.

[22] Ngoc C. Lê, Nguyen The Lam, Son Hong Nguyen, and Duc Thanh Nguyen. 2020. On Vietnamese sentiment analysis: A transfer learning method. In *Proceedings of the RIVF International Conference on Computing and Communication Technologies (RIVF'20)*. IEEE, 1–5. DOI: https://doi.org/10.1109/RIVF48685.2020.9140757

[23] 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.

[24] Long Mai and Bac Le. 2021. Joint sentence and aspect-level sentiment analysis of product comments. *Ann. Oper. Res.* 300, 2 (2021), 493–513.

[25] Yue Mao, Yi Shen, Chao Yu, and Longjun Cai. 2021. A joint training dual-MRC framework for aspect based sentiment analysis. *Proc. AAAI Conf. Artif. Intell.* 35, 15 (May 2021), 13543–13551. DOI: https://doi.org/10.1609/aaai.v35i15.17597

[26] Ambreen Nazir, Yuan Rao, Lianwei Wu, and Ling Sun. 2022. Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. *IEEE Trans. Affect. Comput.* 13, 2 (2022), 845–863. DOI: https://doi.org/10.1109/TAFFC.2020.2970399

[27] Duy Nguyen Ngoc, Tuoi Phan Thi, and Phuc Do. 2021. Preprocessing improves CNN and LSTM in aspect-based sentiment analysis for Vietnamese. In *Proceedings of 5th International Congress on Information and Communication Technology*, Xin-She Yang, R. Simon Sherratt, Nilanjan Dey, and Amit Joshi (Eds.). Springer, 175–185.

[28] Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, 1037–1042. DOI: https://doi.org/10.18653/v1/2020.findings-emnlp.92

[29] Huyen T. M. 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. *J. Comput. Sci. Cyber.* 34, 4 (2018), 295–310.

[30] Minh-Hao Nguyen, Tri Minh Nguyen, Dang Van Thin, and Ngan Luu-Thuy 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)*. IEEE, 1–5. DOI: https://doi.org/10.1109/KSE.2019.8919448

[31] Ngoc Duy Nguyen, Tuoi Phan Thi, and Phuc Do. 2019. A data preprocessing method to classify and summarize aspect-based opinions using deep learning. In *Intelligent Information and Database Systems*, Ngoc Thanh Nguyen, Ford Lumban Gaol, Tzung-Pei Hong, and Bogdan Trawiński (Eds.). Springer International Publishing, Cham, 115–127.

[32] Ngoc-Tu Nguyen, Trong-Dat Nguyen, Duy-Cat Can, Mai-Vu Tran, Ha Luu Manh, and Hoang-Quynh Le. 2021. Attention-based deep learning model for aspect classification on Vietnamese E-commerce data. In *Proceedings of the 13th International Conference on Knowledge and Systems Engineering (KSE'21)*. IEEE, 1–6. DOI: https://doi.org/10.1109/KSE53942.2021.9648690

[33] Ruba Obiedat, Duha Al-Darras, Esra Alzaghoul, and Osama Harfoushi. 2021. Arabic aspect-based sentiment analysis: A systematic literature review. *IEEE Access* 9 (2021), 152628–152645. DOI: https://doi.org/10.1109/ACCESS.2021.3127140

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

[35] Luan-Nghia Pham, Viet-Hong Tran, and Vinh-Van Nguyen. 2013. Vietnamese text accent restoration with statistical machine translation. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC'13)*. Department of English, National Chengchi University, 423–429. Retrieved from https://aclanthology.org/Y13-1044

[36] Long Phan, Hieu Tran, Hieu Nguyen, and Trieu H. Trinh. 2022. ViT5: Pretrained text-to-text transformer for Vietnamese language generation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*. Association for Computational Linguistics, 136–142. DOI: https://doi.org/10.18653/v1/2022.naacl-srw.18

[37] 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

[38] Ton Nu Thi Sau, Do Phuoc Sang, and Pham Thi Thu Trang. 2021. Aspect-based sentiment analysis on student's feedback in Vietnamese. *TNU J. Sci. Technol.* 226, 18 (2021), 48–55.

[39] Samson Tan and Shafiq Joty. 2021. Code-mixing on Sesame Street: Dawn of the adversarial polyglots. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 3596–3616. DOI: https://doi.org/10.18653/v1/2021.naacl-main.282

[40] Kim Nguyen Thi Thanh, Sieu Huynh Khai, Phuc Pham Huynh, Luong Phan Luc, Duc-Vu Nguyen, and Kiet Nguyen Van. 2021. Span detection for aspect-based sentiment analysis in Vietnamese. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*. Association for Computational Lingustics, 318–328.

[41] Dang Van Thin, Duc-Vu Nguyen, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen, and Anh Hoang-Tu Nguyen. 2020. Multi-task learning for aspect and polarity recognition on Vietnamese datasets. In *Computational Linguistics*, Le-Minh Nguyen, Xuan-Hieu Phan, Kôiti Hasida, and Satoshi Tojo (Eds.). Springer, 169–180.

[42] Dang Van Thin, Vu Duc Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy 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. DOI: https://doi.org/10.1109/NICS.2018.8606857

[43] Van Dang Thin, Duong Ngoc Hao, and Ngan Luu-Thuy Nguyen. 2022. An effective contextual language ensemble model for Vietnamese aspect-based sentiment analysis. In *Proceedings of the 9th NAFOSTED Conference on Information and Computer Science (NICS'22)*. IEEE, 30–34. DOI: https://doi.org/10.1109/NICS56915.2022.10013429

[44] Van Dang Thin, Lac Si Le, Hao Minh Nguyen, and Ngan Luu-Thuy Nguyen. 2022. A joint multi-task architecture for document-level aspect-based sentiment analysis in Vietnamese. *Int. J. Mach. Learn. Comput.* 12, 4 (2022), 126–135.

[45] Van Dang 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. DOI: https://doi.org/10.1145/3446678

[46] Van Dang Thin, Vu Duc Nguyen, Nguyen Van Kiet, and Nguyen Luu Thuy Ngan. 2018. A transformation method for aspect-based sentiment analysis. *J. Comput. Sci. Cyber.* 34, 4 (2018), 323–333.

[47] Nguyen Thi Thanh Thuy, Ngo Xuan Bach, and Tu Minh 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. DOI: https://doi.org/10.1109/KSE.2018.8573395

[48] Nguyen Thi Thanh Thuy, Ngo Xuan Bach, and Tu Minh Phuong. 2020. Leveraging foreign language labeled data for aspect-based opinion mining. In *Proceedings of the RIVF International Conference on Computing and Communication Technologies (RIVF'20)*. IEEE, 1–6. DOI: https://doi.org/10.1109/RIVF48685.2020.9140735

[49] Nguyen Luong Tran, Duong Minh Le, and Dat Quoc Nguyen. 2022. BARTpho: Pre-trained sequence-to-sequence models for Vietnamese. In *Proceedings of the 23rd Annual Conference of the International Speech Communication Association*. International Speech Communication Association.

[50] Oanh Thi Tran and Viet The Bui. 2020. A BERT-based hierarchical model for Vietnamese aspect based sentiment analysis. In *Proceedings of the 12th International Conference on Knowledge and Systems Engineering (KSE'20)*. IEEE, 269–274. DOI: https://doi.org/10.1109/KSE50997.2020.9287650

[51] Quang-Linh Tran, Phan Thanh Dat Le, and Trong-Hop Do. 2022. Aspect-based sentiment analysis for Vietnamese reviews about beauty product on E-commerce websites. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation.* 767–776. Retrieved from https://aclanthology.org/2022.paclic-1.84

[52] Tri Cong-Toan Tran, Thien Phu Nguyen, and Thanh-Van Le. 2021. HSUM-HC: Integrating BERT-based hidden aggregation to hierarchical classifier for Vietnamese aspect-based sentiment analysis. In *Proceedings of the 8th NAFOSTED Conference on Information and Computer Science (NICS'21).* IEEE, 284–289. DOI: https://doi.org/10.1109/NICS54270.2021.9701518

[53] Dang Van Thin, Duong Ngoc Hao, Vu Xuan Hoang, and Ngan Luu-Thuy Nguyen. 2022. Investigating monolingual and multilingual BERT models for Vietnamese aspect category detection. In *Proceedings of the RIVF International Conference on Computing and Communication Technologies (RIVF'22).* IEEE, 130–135. DOI: https://doi.org/10.1109/RIVF55975.2022.10013792

[54] Yifei Yang and Hai Zhao. 2022. Aspect-based sentiment analysis as machine reading comprehension. In *Proceedings of the 29th International Conference on Computational Linguistics.* International Committee on Computational Linguistics, 2461–2471. Retrieved from https://aclanthology.org/2022.coling-1.217

[55] Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021. Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, 9209–9219. DOI: https://doi.org/10.18653/v1/2021.emnlp-main.726

[56] Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A Survey on Aspect-based Sentiment Analysis: Tasks, Methods, and Challenges. DOI: https://doi.org/10.48550/ARXIV.2203.01054

Received 6 February 2023; revised 22 May 2023; accepted 13 July 2023