Machine Learning and Data Mining

Lecturer: Dr. Nguyen Duc Anh

Class ID: 156818

Aspect-based Sentiment Analysis for Vietnamese

AUTHORS

Vu Hai Dang - 20225962
Nguyen Minh Khoi - 20226050
Le Dai Lam - 20225982
Pham Thanh Nam - 20225989
Bui Hoang Viet - 20226073

Contents

1 Introduction

Nowadays, the popularity of the Internet has led to an explosion of enormous data sources from users, especially in fields such as e-commerce, social networks, and search engines. This has brought about continuous Artificial Intelligence (AI) development, especially Natural Language Processing (NLP). Among them, the problem of Sentiment Analysis is increasingly popular and successful in both research and commerce with the purpose is to understand the level of satisfaction through reviews of customers. However, it has not fully exploited the valuable data from the Internet because a user review can contain helpful information for companies or research institutions. Aspect-based Sentiment Analysis (ABSA) is an improvement of Sentiment Analysis that will solve that problem.

Aspect-based Sentiment Analysis is a technique of analyzing text, classifying data by aspect, and identifying sentiment polarities for each of those aspects. For example, in the Hotel domain, aspects might be the experience of customers about service, the response time of hotel complaints, or quality of room amenities. With the popularity of the Internet, smart mobile, everyone can now easily evaluate aspects of a product or a service quickly and conveniently.

The huge review response data source is a valuable resource for companies. Companies collect and extract valuable information to understand what customers want and need to drive growth. The problem of Aspect-based Sentiment Analysis is the key point to extract that information. In the field of Natural Language Processing, the ABSA problem was first proposed at SemEval-2014 Task 4 [8] by Pontiki et al. After that, there were many big competitions on analyzing the sentiment aspect were organized and achieved outstanding results such as SemEval 2015 task 12 [7], SemEval 2016 task 5 [6]. In addition, with the appearance of large, powerful language models such as BERT [1], Natural Language Processing problems have made great strides.

This article experiments with powerful, modern deep learning models for the ABSA problem on the VLSP 2018 ABSA dataset with two domains: Restaurants and Hotels. The ABSA problem that we performed determines two problems: Aspect Category Detection (ACD) and Sentiment Polarity Classification (SPC). The metric we use for the evaluation is the average micro F1-score.

In this article, we focus on introducing information related to the problem of Aspect-based Sentiment Analysis in Vietnamese language.

2 Problem dessription

Specifically, the problem we perform is **Aspect Category Sentiment Analysis (ACSA)**, and we divide this problem into two subproblems: **Aspect Category Detection** and **Sentiment Polarity Classification** based on detected Aspect:

2.1 Aspect Category Detection

Identify the entity **E** and attribute **A** of a review expressed in a given sentence. **E** and **A** should be selected from a predefined set of entity types (e.g. "ROOMS","HOTEL") and attribute labels (e.g. "PRICE","QUALITY").

2.2 Sentiment Polarity Classification

Each identified E#A pair must be assigned one of the following sentiment polarization labels: "Positive", "Negative", "Neutral".

3 Related work

3.1 BERT

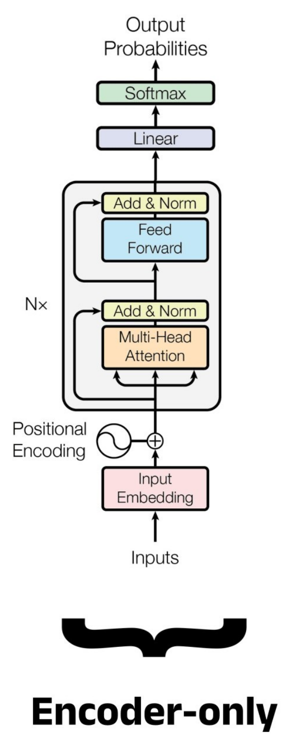

Figure 1: Bidirectional Encoder Representations from Transformers (BERT)

Bidirectional Encoder Representations from Transformers (BERT) is a groundbreaking pre-training technique for Natural Language Processing (NLP) developed by Google in 2018. Before

BERT, many state-of-the-art NLP models relied on pre-trained word embeddings like Word2Vec or GloVe, which captured word meanings but struggled with context-dependent semantics. **BERT** revolutionized the field by introducing a truly bidirectional approach to understanding language, allowing models to grasp the full context of a word by simultaneously considering the words that precede and follow it.

At its core, **BERT** is a large-scale transformer-based model. The Transformer architecture, introduced in 2017, relies heavily on a mechanism called "self-attention," which allows the model to weigh the importance of different words in a sentence when processing a particular word. Unlike previous models that processed text sequentially (e.g., LSTMs or GRUs) or only bidirectionally in a shallow way (e.g., ELMo, which concatenated two separate unidirectional LSTMs), **BERT's multi-headed self-attention mechanism** processes the entire input sequence at once. This parallel processing capability, combined with its deep architecture, enables **BERT** to learn rich, contextualized representations of words.

The "pre-training" aspect is crucial to **BERT**'s success. Instead of training a model from scratch for every new NLP task, BERT is first pre-trained on massive amounts of unlabeled text data, such as the entire English Wikipedia and BookCorpus. This pre-training involves two novel, unsupervised tasks:

- **Masked Language Model (MLM):** In this task, **BERT** randomly masks 15% of the words in a sentence and then tries to predict the original masked words based on the context provided by the unmasked words. This forces the model to learn deep bidirectional representations, as it must understand the context from both left and right to accurately predict the missing words. For example, in the sentence "The man [MASK] to the store," BERT learns to predict "went" by looking at "man," "to," and "store" simultaneously. This is a significant departure from traditional language models that only predict the next word in a sequence.

- **Next Sentence Prediction (NSP):** This task helps **BERT** understand the relationships between sentences. Given two sentences, A and B, **BERT** is trained to predict whether sentence B logically follows sentence A in the original document. This is vital for tasks like question answering and natural language inference, where understanding inter-sentence coherence is critical.

After this extensive pre-training phase, BERT produces a highly versatile language understanding model. The beauty of BERT lies in its "fine-tuning" capability. For specific downstream NLP tasks, such as sentiment analysis, named entity recognition, or question answering, the pre-trained BERT model can be adapted with minimal additional training. This involves adding a small, task-specific output layer on top of the pre-trained BERT model and then training the entire model (BERT's layers plus the new output layer) on a labeled dataset for that specific task. Because BERT has already learned a vast amount of linguistic knowledge during pre-training, fine-tuning typically requires much less labeled data and computational resources compared to training a model from scratch, while often achieving superior performance.

BERT's impact on NLP has been profound, setting new benchmarks across numerous tasks and paving the way for a new era of large-scale pre-trained language models like GPT-3, RoBERTa, and XLNet. Its ability to capture deep contextual relationships and its efficient fine-tuning paradigm have made it an indispensable tool for researchers and practitioners alike, significantly advancing the state of the art in natural language understanding.

3.2 RoBERTa

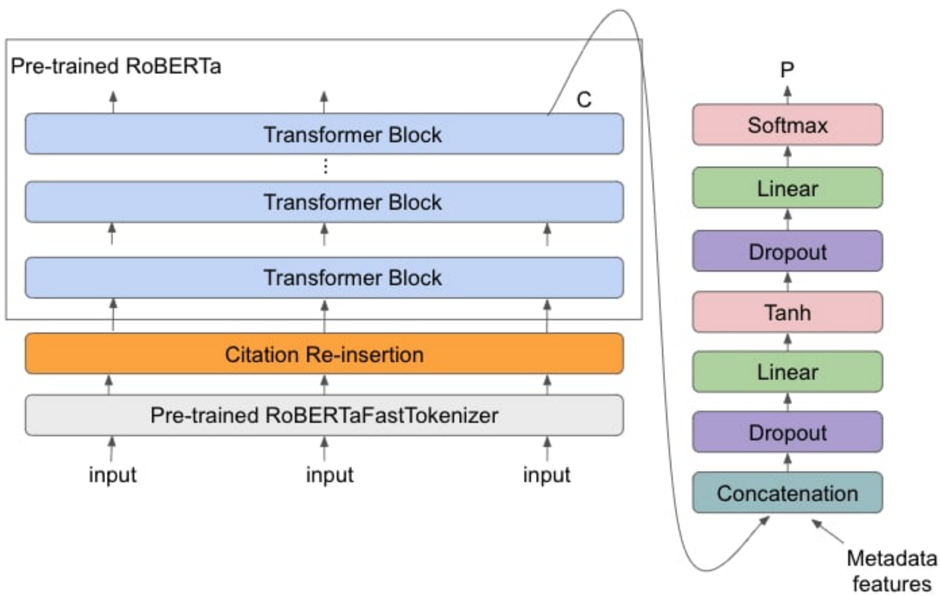

Figure 2: RoBERTa model

RoBERTa, short for **R**obustly **O**ptimized **BERT A**pproach, is a pre-training method for Natural Language Processing (NLP) models that significantly improved upon the original BERT (Bidirectional Encoder Representations from Transformers) model. Developed by Facebook AI in 2019, RoBERTa demonstrated that BERT's performance could be substantially boosted by simply training it longer, with more data, on larger batches, and with a few key modifications to the pre-training objective. It solidified the paradigm of large-scale pre-trained language models as the standard for achieving state-of-the-art results across a wide array of NLP tasks.

The motivation behind RoBERTa was to investigate whether the original BERT model was under-tuned and to rigorously evaluate the impact of various hyperparameter choices and training strategies on its performance. While BERT was revolutionary, the RoBERTa team hypothesized that its full potential had not yet been realized. Their systematic study led to several crucial optimizations:

1. **More Data:** RoBERTa was pre-trained on a significantly larger corpus of text compared to BERT. While BERT used BookCorpus and English Wikipedia (around 16GB of text), RoBERTa utilized a massive 160GB of text data, including CC-News, OpenWebText, and stories from the Common Crawl. This vast increase in exposure to diverse linguistic patterns allowed the model to learn more robust and generalized representations.

2. **Longer Training and Larger Batches:** The original BERT was trained for a fixed number of steps with a relatively small batch size. RoBERTa was trained for much longer (more steps) and with significantly larger batch sizes (up to 8000 sequences per batch). This combination allows the model to see more data points and update its weights more effectively, leading to better convergence and performance.

3. **Dynamic Masking:** BERT used static masking, meaning the same random tokens were masked for prediction in every epoch. RoBERTa introduced dynamic masking, where the masking pattern is generated anew for each training example every time it is fed into the model. This ensures that the model sees a wider variety of masked tokens and contexts over the course of training, preventing it from overfitting to a fixed set of masked positions.

4. **Removal of the Next Sentence Prediction (NSP) Objective:** The original BERT included an NSP task, where the model predicted whether two sentences were consecutive in the original document. The RoBERTa team found that removing this objective, and instead training the model on full sentences or contiguous segments of text, did not hurt performance on downstream tasks and often led to improvements. This simplified the pre-training process and focused the model's learning on capturing within-document coherence through the Masked Language Model (MLM) task alone.

5. **Byte-Pair Encoding (BPE) Tokenization:** Like BERT, RoBERTa uses BPE for tokenization, which effectively handles out-of-vocabulary words by breaking them down into subword units. This allows the model to process a vast vocabulary while keeping the actual vocabulary size manageable.

The cumulative effect of these optimizations was a model that consistently outperformed BERT on a wide range of NLP benchmarks, including GLUE (General Language Understanding Evaluation) and SQuAD (Stanford Question Answering Dataset). RoBERTa demonstrated that simply scaling up and refining the pre-training process of the Transformer architecture could yield substantial gains without introducing fundamentally new architectural components. It underscored the importance of meticulous engineering and extensive computational resources in the era of large language models, setting a new standard for subsequent research and development in the field.

3.3 PhoBERT

Following the success of Bidirectional Encoder Representations from Transformers (BERT) in English Natural Language Processing (NLP), the need for high-performing, language-specific models became evident. This led to the development of **PhoBERT**, a state-of-the-art pre-trained language model specifically designed for the Vietnamese language. Released by VinAI Research, **PhoBERT** is essentially a Vietnamese version of BERT, leveraging the same powerful Transformer architecture but trained exclusively on a massive Vietnamese corpus.

The core motivation behind creating **PhoBERT** was the observation that general-purpose multilingual models, while capable of handling Vietnamese, often underperformed compared to models trained on a single language. This is because language-specific nuances, morphology, syntax, and vocabulary are best captured when a model is exposed to a vast amount of text in that particular language. Vietnamese, being an analytic language with a rich tonal system and complex word segmentation rules (especially compound words), benefits significantly from a dedicated language model.

PhoBERT's architecture is based on RoBERTa (A Robustly Optimized BERT Pretraining Approach), which is an optimized version of BERT. RoBERTa improved upon BERT's pre-training methodology by using more data, training for longer, removing the Next Sentence Prediction (NSP) objective, and dynamically changing the masking pattern. PhoBERT adopts these enhancements, making it even more robust and effective for Vietnamese.

The pre-training of **PhoBERT** is a critical aspect of its performance. It was trained on an enormous dataset of Vietnamese text, comprising approximately 20-35GB of raw text data. This dataset includes a diverse range of sources such as Vietnamese Wikipedia, news articles, and web crawls. Similar to BERT and RoBERTa, **PhoBERT** undergoes unsupervised pre-training using the Masked Language Model (MLM) objective. During this process, random tokens in the Vietnamese text are masked, and the model is tasked with predicting these masked tokens based on their surrounding context. This forces PhoBERT to learn deep, bidirectional representations of Vietnamese words and phrases, capturing intricate linguistic patterns and semantic relationships unique to the language.

The impact of **PhoBERT** on Vietnamese NLP has been transformative. Before its advent, researchers and developers often had to rely on less effective methods or adapt English models, which often yielded suboptimal results due to the significant linguistic differences. **PhoBERT** has consistently achieved new state-of-the-art results across a wide array of Vietnamese NLP tasks, including:

- Sentiment Analysis: Understanding the emotional tone of Vietnamese text.
- Named Entity Recognition (NER): Identifying and classifying named entities (e.g., person names, locations, organizations) in Vietnamese sentences.
- Text Classification: Categorizing Vietnamese documents or sentences into predefined classes.
- Question Answering: Extracting answers from Vietnamese texts based on given questions.

By providing a powerful, pre-trained foundation, PhoBERT significantly reduces the computational resources and labeled data required for fine-tuning on specific downstream tasks. This has democratized advanced NLP capabilities for Vietnamese, enabling faster development of more accurate and robust applications. PhoBERT stands as a testament to the importance of language-specific models in achieving true linguistic understanding.

4 Dataset

4.1 The VLSP 2018 Aspect-based Sentiment Analysis Dataset

The VLSP 2018 Aspect-based Sentiment Analysis (ABSA) Dataset stands as a pivotal resource for advancing Natural Language Processing (NLP) research in Vietnamese, particularly in the nuanced domain of sentiment analysis. Developed as part of the Vietnam Language and Speech Processing (VLSP) workshop's shared task in 2018, this dataset provided a standardized benchmark for evaluating models capable of understanding fine-grained sentiment in Vietnamese text.

Traditional sentiment analysis typically classifies the overall sentiment of a document or sentence (positive, negative, neutral). However, Aspect-based Sentiment Analysis goes a step further by identifying the specific aspects or features of an entity mentioned in a text and determining the sentiment expressed towards each of those aspects. For instance, in a hotel review, a sentence might express positive sentiment towards the "room cleanliness" but negative sentiment towards "hotel price." The VLSP 2018 dataset was specifically designed to tackle this challenge for Vietnamese.

The dataset primarily focuses on **hotel reviews**, a common and rich source of aspect-level sentiment. The core task defined for this dataset is **Aspect Category Sentiment Analysis (ACSA)**. This involves two main sub-problems:

1. **Aspect Category Detection:** Identifying the entity (E) and attribute (A) pairs (e.g., "ROOMS#DESIGN", "SERVICE#GENERAL") expressed in a given sentence. These entities and attributes are selected from a predefined set of types relevant to the hotel domain.

2. **Sentiment Polarity Classification:** Assigning one of three sentiment polarization labels ("Positive", "Negative", "Neutral") to each detected E#A pair.

- {cate: "ROOMS#DESIGN", pol: "positive"}

- {cate: "SERVICE#GENERAL", pol: "positive"}

The dataset is typically split into training, development (dev), and test sets, allowing researchers to train their models, tune hyperparameters, and evaluate performance on unseen data. The training set is the largest, providing ample examples for model learning, while the dev set helps in refining the model, and the test set offers an unbiased evaluation of generalization capabilities. These splits are crucial for ensuring fair comparisons between different ABSA models.

The significance of the VLSP 2018 ABSA dataset cannot be overstated. Before its release, standardized, high-quality annotated datasets for Vietnamese ABSA were scarce. This dataset filled a critical gap, providing a common ground for researchers to develop and test their algorithms. It spurred significant advancements in Vietnamese NLP, encouraging the community to explore various machine learning and deep learning approaches tailored to the complexities of the Vietnamese language, such as its rich morphology, lack of explicit word boundaries, and unique grammatical structures. The dataset highlighted the challenges of robust word segmentation and accurate aspect extraction in Vietnamese.

In essence, the VLSP 2018 ABSA dataset served as a foundational resource, pushing the boundaries of aspect-based sentiment analysis for Vietnamese and contributing significantly to the broader landscape of Vietnamese NLP research. Its impact continues to be felt as a benchmark for subsequent work in fine-grained sentiment analysis.

Domain	**Dataset**	**No. Reviews**	**No. Aspect #**	**Avg. Length**	**Vocab Size**	**No. words in Test / Dev not in Training set**
Hotel	Training	3,000	13,948	47	3,908	-
	Dev	2,000	7,111	23	2,745	1,059
	Test	600	2,584	30	1,631	346

Table 1: Dataset Statistics for Hotel Domain

The Hotel domain consists of 34 following Aspect#Category pairs:

- **Facilities**
  - facilities#cleanliness
  - facilities#comfort
  - facilities#design&features
  - facilities#general
  - facilities#miscellaneous
  - facilities#prices
  - facilities#quality

- **Food&Drinks**
  - food&drinks#miscellaneous
  - food&drinks#prices
  - food&drinks#quality
  - food&drinks#style&options
- **Hotel**
  - hotel#cleanliness
  - hotel#comfort
  - hotel#design&features
  - hotel#general
  - hotel#miscellaneous
  - hotel#prices
  - hotel#quality
- **Location**
  - location#general
- **Rooms**
  - rooms#cleanliness
  - rooms#comfort
  - rooms#design&features
  - rooms#general
  - rooms#miscellaneous
  - rooms#prices
  - rooms#quality
- **Room_Amenities**
  - room_amenities#cleanliness
  - room_amenities#comfort
  - room_amenities#design&features
  - room_amenities#general

  - room_amenities#miscellaneous
  - room_amenities#prices
  - room_amenities#quality

- **Service**
  - service#general

4.2 Data preprocessing

Natural language processing in general and Vietnamese processing in particular, the first step is to preprocess the data. If this step is done well, it can increase performance. Therefore, data preprocessing is the process of normalizing data and removing non-significant components. So we designed a pipeline to process the data in order to get better information from the original data. These steps are described as below:

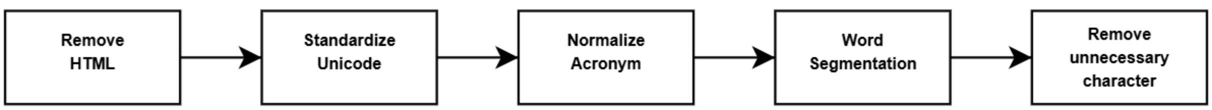

Figure 3: Preprocessing Pipeline

First, we deleted the HTML codes in the original dataset. We then performed the charset standardization from Windows1252 to UTF-8 and standardized common Vietnamese abbreviations for each domain in the dataset. In Vietnamese, a word can be made up of two or more other words (known as a compound word, each word has its meaning when standing alone and another meaning when combined). Therefore, it is necessary to do the segmentation for Vietnamese words before starting further processing, and we used the VnCoreNLP toolkit [13] to implement that step. Finally, we removed unnecessary characters, this helps to reduce the number of feature dimensions for sentence, increase the speed, and avoids bad influence on the model results.

We implemented 3 classes in `vietnamese_processor.py` to preprocess raw Vietnamese text data. This is my improved version from the work by behitek:

(a) **VietnameseTextCleaner**: Simple regex-based text cleaning to remove HTML, Emoji, URL, Email, Phone Number, Hashtags, and other unnecessary characters.

(b) **VietnameseToneNormalizer**: Normalize Unicode.

(c) **VietnameseTextPreprocessor**: Combines the above classes and adds these following steps to the pipeline:

- `normalize_teencodes(text:str)`: Converts teencodes to their original form. I also provided the `extra_teencodes` parameter to add your own teencode definitions based on the dataset used. The `extra_teencodes` must be a `dict` with keys as the original form and values as a list of teencodes. You should be careful when using single word replacement for teencodes, because it can cause misinterpretation.
- `correct_vietnamese_errors(texts:List)`: Uses pre-trained model by bmd1905 to correct Vietnamese errors. The inference time for this model is quite slow, so I implemented this method to process the text in batch. That's why you should pass a `list of texts` as input.
- `word_segment(text:str)`: Use VnCoreNLP to segment Vietnamese words. This tool is chosen because: "PhoBERT employed the RDRSegmenter from VnCoreNLP to pre-process the pretraining data". I already implemented a script to automatically download necessary components of this tool into the `VnCoreNLP` folder, so you don't need to do anything.

5 Model Architecture

According to the original BERT paper, the Feature Extraction Strategy that concatenates the last four hidden layers of BERT yields the best performance. We applied that method to the PhoBERT layer in our model architectures and combined it with two output construction methods, ACSA-v1 and ACSA-v2, to form our final solutions.

5.1 ACSA-v1: Multi-task Approach

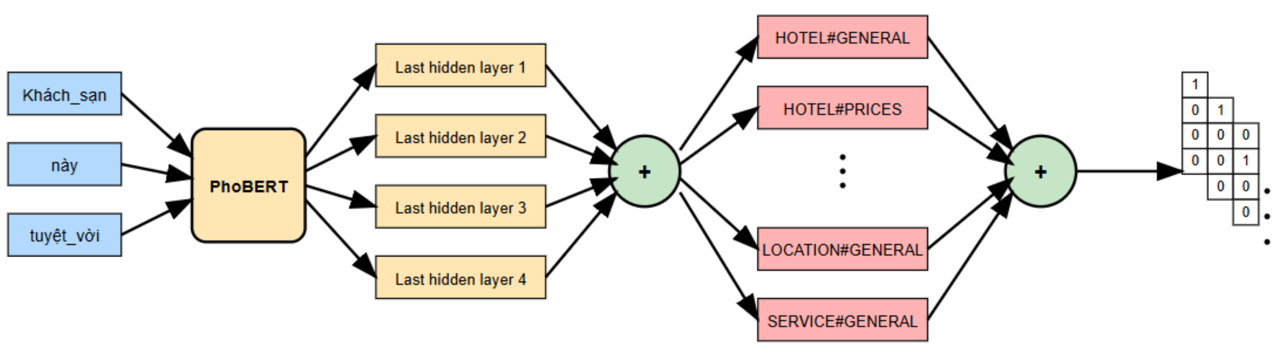

Figure 4: ACSA-v1: Multi-task Approach

5.1.1 Output Construction

We transformed each Aspect#Category pair and their corresponding Polarity labels in each dataset's review into a list of C one-hot vectors, where C is the number of Aspect#Category pairs.

Each vector has 4 components:

- Positive
- Negative
- Neutral
- None (indicating whether the Aspect#Category exists in the input)

If a label exists, its corresponding component is set to 1; otherwise, it is 0.

Therefore, we create C Dense layers, each with 4 neurons to predict the polarity of the corresponding Aspect#Category pair. A softmax function is applied to each 4-neuron group to get the probability distribution over the 4 polarity classes.

Rather than feeding the learned features to each Dense layer separately, we concatenate them into a single Dense layer with:

- $34 \times 4 = 136$ neurons for the Hotel domain
- $12 \times 4 = 48$ neurons for the Restaurant domain

We apply the `binary_crossentropy` loss function to treat each 4-neuron group as a binary classification problem.

5.1.2 Why Use One-hot Encoding and Softmax?

In this ACSA task, each `Aspect#Category,Polarity` can be viewed as an independent binary classification (e.g., "Is this Aspect#Category Positive? Negative?").

However, using Sigmoid on each label independently could lead to conflicting outputs, such as:

$$\text{Positive} = 0.9, \quad \text{Negative} = 0.8, \quad \text{Neutral} = 0.7$$

This contradicts the nature of the task, as these polarities are mutually exclusive. Softmax ensures the output probabilities sum to 1, better capturing the semantic relationship among polarity labels.

5.1.3 Why Concatenate into a Single Dense Layer and Use Binary Crossentropy?

Concatenation mixes independent `Aspect#Category,Polarity` representations and enables the model to learn complex, shared relationships between them. For instance, if `HOTEL#CLEANLINESS` is Positive, the model might infer `HOTEL#QUALITY` is likely to be Positive too.

By using concatenation:

- The Softmax constraint is maintained for each 4-neuron polarity group.
- `binary_cross_entropy` is applied independently to each of these groups.
- The model learns to predict all `Aspect#Category,Polarity` instances simultaneously.

5.2 ACSA-v2: Multi-task with Multi-branch Approach

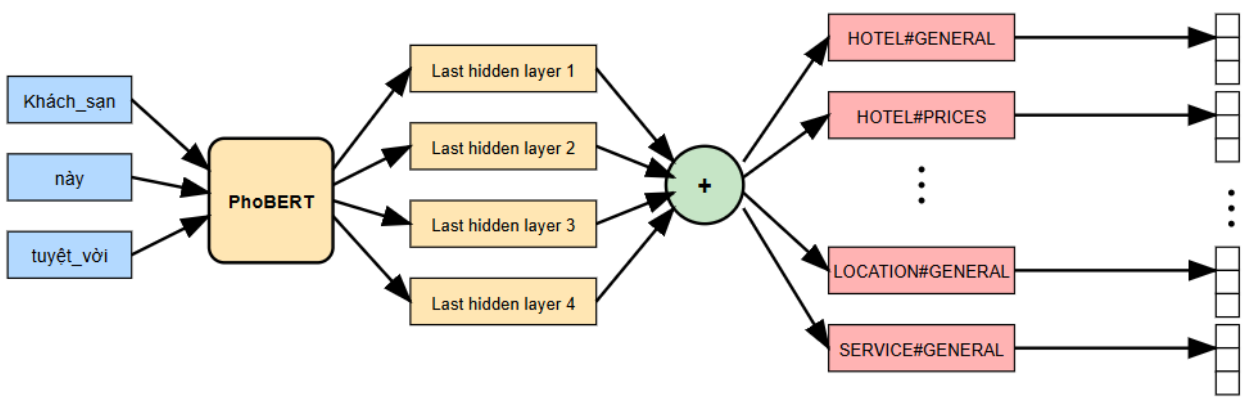

Figure 5: ACSA-v2: Multi-task with Multi-branch Approach

This approach differs from ACSA-v1 in that it branches into multiple sub-models using C Dense layers (34 for Hotel, 12 for Restaurant), without concatenating them into one.

Each Dense layer predicts its corresponding task independently. The Softmax function is applied directly to each Dense layer output (with 4 units) to produce the polarity distribution.

`categorical_cross_entropy` loss is used to treat each Dense layer as a multi-class classification problem.

6 Experiment & Result

6.1 Experimental Setup

The experiments were conducted on the VLSP 2018 ABSA dataset, specifically focusing on the Hotel domain. The dataset comprises:

- **Training set**: 3,000 reviews with 13,948 `Aspect#Category, Polarity` annotations.
- **Development set**: 2,000 reviews with 7,111 annotations.

- **Test set**: 600 reviews with 2,584 annotations.

The Hotel domain includes 34 distinct `Aspect#Category` pairs. For model training and evaluation, the following configurations were employed:

- **Pre-trained model**: PhoBERT.
- **Input sequence length**: 256 tokens.
- **Batch size**: 20.
- **Number of epochs**: 20, with early stopping applied.
- **Optimizer**: AdamW.
- **Learning rate schedule**: Linear decay with a warm-up phase covering 10% of the total steps.
- **Initial learning rate**: 2×10^{-4}.

6.2 Results on the Hotel Domain

The models were evaluated on two tasks:

- **Aspect Category Detection (ACD)**: Identifying the presence of specific `Aspect#Category` pairs in reviews.
- **Aspect Category Detection with Sentiment Polarity Classification (ACD + SPC)**: Identifying `Aspect#Category` pairs along with their associated sentiment polarity (Positive, Negative, Neutral).

The evaluation metric used was the micro F1-score. The results are presented in Table **??**.

6.3 Discussion

The **Multi-task** model achieved the highest F1-scores in both tasks, outperforming previous state-of-the-art methods. This indicates the effectiveness of the proposed approach in jointly learning aspect detection and sentiment classification. The **Multi-task Multi-branch** model, while conceptually modular, showed lower performance, suggesting that shared representations in a unified model are more beneficial for this problem.

Task	Method	Precision	Recall	F1-score
Aspect# Category	VLSP best submission	76.00	66.00	70.00
	Bi-LSTM+CNN	84.03	72.52	77.85
	BERT-based Hierarchical	-	-	82.06
	Multi-task	87.45	78.17	82.55
	Multi-task Multi-branch	63.21	57.86	60.42
Aspect# Category, Polarity	VLSP best submission	66.00	57.00	61.00
	Bi-LSTM+CNN	76.53	66.04	70.90
	BERT-based Hierarchical	-	-	74.69
	Multi-task	81.90	73.22	77.32
	Multi-task Multi-branch	57.55	52.67	55.00

Table 2: Performance Comparison of Methods on Aspect# Category and Aspect# Category, Polarity Tasks

6.4 Conclusion

The experimental results on the VLSP 2018 ABSA Hotel dataset highlight the effectiveness of our proposed Multi-task learning approach based on PhoBERT. Specifically, the model achieved the highest performance across two main tasks: Aspect Category Detection (ACD) with an F1-score of **82.55%**, and ACD with Sentiment Polarity Classification (ACD + SPC) with an F1-score of **77.32%**. These results surpass previous baselines such as Bi-LSTM + CNN and BERT-based Hierarchical models.

The success of the Multi-task model stems from its ability to jointly learn shared representations for aspect and polarity classification. In contrast, the Multi-task Multi-branch variant, which predicts each aspect independently, yielded notably lower performance. This supports the hypothesis that sharing parameters across related tasks helps the model better capture contextual and semantic dependencies.

In summary, the combination of PhoBERT and multi-task learning not only improves the accuracy of aspect-level sentiment analysis but also demonstrates strong generalizability across fine-grained classification tasks. These findings reinforce the importance of leveraging pre-trained language models and multi-task objectives in resource-scarce languages like Vietnamese.

7 Improvement

To enhance the model's generalization and robustness, we independently implemented a data augmentation strategy based on **back-translation**, using MarianMT models. This method involves translating a Vietnamese review into English and then translating it back to Vietnamese to produce a semantically equivalent but syntactically varied version. Such augmentation helps increase lexical diversity while preserving the original sentiment and partially address data imbalance.

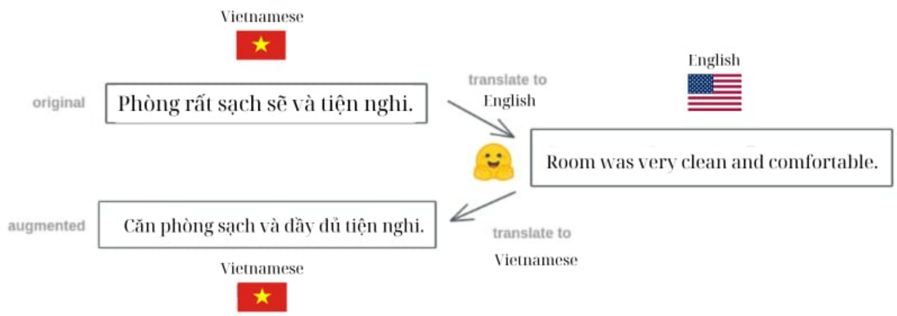

Figure 6: An example of back-translation for data augmentation.

The process was conducted as follows:

- **Models used:** `Helsinki-NLP/opus-mt-vi-en` and `Helsinki-NLP/opus-mt-en-vi` from HuggingFace's Transformers library.
- **Sample selection:** Approximately 7% (1 out of every 15 samples) of the training dataset was randomly selected for augmentation.
- **Procedure:**
  - Each selected review was first translated from Vietnamese to English.
  - The English translation was then translated back into Vietnamese.
  - The resulting back-translated review was added to the training dataset while preserving the original aspect-polarity labels.

After applying this augmentation method, the training dataset increased from 3,000 to approximately 3,200 reviews, resulting in a richer and more varied input space.

Our experiments showed that this augmentation approach **improved the model's F1 score by more than 1% from 0.7255 to 0.7378)** , particularly benefiting the classification of rare aspect-polarity combinations.

These results highlight the effectiveness of lightweight, language-agnostic augmentation techniques in resource-scarce languages like Vietnamese and their practical impact on enhancing model performance without architectural changes.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[2] H. H. Do, P. Prasad, A. Maag, and A. Alsadoon, "Deep learning for aspect-based sentiment analysis: a comparative review," *Expert Systems with Applications*, vol. 118, pp. 272–299, 2019.

[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[4] D. Q. Nguyen and A. T. Nguyen, "Phobert: Pre-trained language models for vietnamese," *arXiv preprint arXiv:2003.00744*, 2020.

[5] H. T. Nguyen, H. V. Nguyen, Q. T. Ngo, L. X. Vu, V. M. Tran, B. X. Ngo, and C. A. Le, "Vlsp shared task: sentiment analysis," *Journal of Computer Science and Cybernetics*, vol. 34, no. 4, pp. 295–310, 2018.

[6] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. AL-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, V. Hoste, M. Apidianaki, X. Tannier, N. Loukachevitch, E. Kotelnikov, N. Bel, S. M. Jiménez-Zafra, and G. Eryiğit, "Semeval-2016 task 5: Aspect based sentiment analysis," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, (San Diego, California), pp. 19–30, Association for Computational Linguistics, 2016.

[7] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos, "Semeval-2015 task 12: Aspect based sentiment analysis," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, (Denver, Colorado), pp. 486–495, Association for Computational Linguistics, 2015.

[8] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "Semeval-2014 task 4: Aspect based sentiment analysis," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, (Dublin, Ireland), pp. 27–35, Association for Computational Linguistics, 2014.

[9] S. Ruder, P. Ghaffari, and J. G. Breslin, "A hierarchical model of reviews for aspect-based sentiment analysis," *arXiv preprint arXiv:1609.02745*, 2016.

[10] T. T. Thet, J.-C. Na, and C. S. Khoo, "Aspect-based sentiment analysis of movie reviews on discussion boards," *Journal of of information science*, vol. 36, no. 6, pp. 823–848, 2010.

[11] O. T. Tran and V. T. Bui, “A bert-based hierarchical model for vietnamese aspect based sentiment analysis,” in *2020 12th International Conference on Knowledge and Systems Engineering (KSE)*, pp. 269–274, 2020.

[12] D. V. Thin, D.-V. Nguyen, K. V. Nguyen, N. L.-T. Nguyen, and A. H.-T. Nguyen, “Multi-task learning for aspect and polarity recognition on vietnamese datasets,” in *International Conference of the Pacific Association for Computational Linguistics*, pp. 169–180, Springer, 2019.

[13] T. Vu, D. Q. Nguyen, D. Q. Nguyen, M. Dras, and M. Johnson, “Vncorenlp: A vietnamese natural language processing toolkit,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, (New Orleans, Louisiana), pp. 56–60, Association for Computational Linguistics, 2018.

[14] W. Xue and T. Li, “Aspect based sentiment analysis with gated convolutional networks,” *arXiv preprint arXiv:1805.07043*, 2018.