

Project Proposal

Aspect-Based Sentiment Analysis

Can Ha An, Nguyen Mau Hoang Hiep, Nguyen Xuan Thai Duong

April 22, 2025

1. Problem Description and Significance

Aspect-Based Sentiment Analysis (ABSA) is a task in natural language processing that identifies specific aspects mentioned in a text and determines the sentiment expressed toward each aspect. For example, in the sentence "The food was great but the service was slow," ABSA would identify two aspects—food and service—and classify the sentiment toward each as positive and negative.

This project explores ABSA in Vietnamese using the VLSP 2018 ABSA dataset. Our goal is to develop a strong transformer-based model using PhoBERT for this task, skipping over traditional machine learning models which do not leverage contextual embeddings effectively.

The significance of this work lies in developing sentiment analysis tools that can better handle language-specific challenges and cultural contexts. By improving ABSA in Vietnamese, we also help extend NLP research to underrepresented languages. This has practical applications in areas like customer review analysis and broader importance for building inclusive, multilingual NLP systems.

2. Dataset Details

VLSP 2018 ABSA (Vietnamese)

- **Source:** Vietnamese Language and Speech Processing (VLSP) 2018 [1]
- **Structure:** The VLSP 2018 ABSA dataset includes annotated customer reviews from two domains: restaurants and hotels, split into training, development, and testing sets. Each review is labeled with one or more aspect categories, represented as **entity#attribute** pairs like **food#quality**, **service#style**, and **hotel#prices**. Each aspect is annotated with a sentiment label (Positive, Negative, or Neutral).

Domain	Train	Dev	Test	Aspect Categories
Restaurant	2,961	1,290	500	6 entities \times 5 attributes = 12 pairs
Hotel	3,000	1,200	600	7 entities \times 8 attributes = 34 pairs

3. Benchmark Algorithm and Model Development

Rather than benchmarking too extensively across legacy models, we will go straight to implementing a state-of-the-art PhoBERT model with a classification head. PhoBERT is a monolingual Vietnamese language model based on RoBERTa and trained on a large-scale Vietnamese corpus [?]. We will explore various architecture choices for the classification head, such as concatenating aspect and sentence as input or using aspect-specific attention mechanisms.

As a lightweight benchmark, we also include a Support Vector Machine (SVM) classifier using TF-IDF features to provide a traditional baseline. This helps contextualize the performance gains from using transformer-based models. Previous work using an SVM-based method reported an F1

score of 0.70 on the VLSP2018 hotel dataset [2]. Additionally, a CNN-based approach achieved an F1 of 0.80 on the restaurant dataset [3]. These results provide reference points for minimum expected model performance.

4. Evaluation Metrics

We evaluate our models using the following metrics: **F1-Score (Macro Average)** for aspect category detection, and **Accuracy**, **Precision**, and **Recall** for sentiment classification.

5. Project Workflow

1. Data Collection and Preprocessing

- Handle missing values, remove noise, and normalize data.
- Apply text preprocessing techniques such as tokenization, stopword removal, and stemming.
- Generate word embeddings (e.g., Word2Vec, GloVe, FastText, etc.)

2. Model Development

- Load PhoBERT and add a classification head.
- Experiment with different input formats and head designs.
- Train a baseline SVM model with TF-IDF features.
- Fine-tune PhoBERT-based classifiers using training and dev sets.

3. Model Evaluation and Reporting

- Generate predicted labels and compare against ground-truth labels
- Evaluate different variations of PhoBERT on the test set using selected metrics
- Compare performance between SVM and PhoBERT
- Visualize confusion matrix, per-aspect performance, etc.

4. Documentation and Reporting

- Document approach, observations, and performance.
- Prepare results for reporting or deployment.

References

- [1] Absa vietnamese vlsp2018 datasets. <https://github.com/MLlove0402/ABSA-Vietnameses-VLSP2018-Datasets/tree/master/VLSP2018>, 2025. Accessed: 2025-04-04.
- [2] Huyen T M Nguyen, Hung V Nguyen, Quyen T Ngo, Luong X Vu, Vu Mai Tran, Bach X Ngo, and Cuong A Le. Vlsr shared task: Sentiment analysis. *Journal of Computer Science and Cybernetics*, 34(4):295–310, Jan. 2019.
- [3] Dang Van Thin, Vu Duc Nguye, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. Deep learning for aspect detection on vietnamese reviews. In *2018 5th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 104–109, 2018.