VIETNAM GENERAL CONFEDERATION OF LABOR
TON DUC THANG UNIVERSITY
FACULTY OF INFORMATION TECHNOLOGY

HOANG LE MINH NHAT
PHAM NGUYEN ANH
NGUYEN QUANG HUY

FINAL REPORT

KNOWLEDGE DISCOVERY AND DATA MINING

HO CHI MINH CITY, 2025

VIETNAM GENERAL CONFEDERATION OF LABOR

TON DUC THANG UNIVERSITY

FACULTY OF INFORMATION TECHNOLOGY

HOANG LE MINH NHAT - 523H0064

PHAM NGUYEN ANH - 523H0117

NGUYEN QUANG HUY - 523H0140

FINAL REPORT

KNOWLEDGE DISCOVERY AND DATA MINING

Advised by

Dr. Hoang Anh

HO CHI MINH CITY, 2025

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to Dr. Hoang Anh, our instructor and mentor, for his valuable guidance and support throughout the final report. He has been very helpful and patient in providing us with constructive feedback and suggestions to improve our work. He has also encouraged us to explore new technologies and techniques to enhance our system's functionality and performance. We have learned a lot from his expertise and experience in web development and software engineering. We are honored and privileged to have him as our teacher and supervisor.

Ho Chi Minh city, 23th December 2025.

Author

(Signature and full name)

Nhat

Hoang Le Minh Nhat

Anh

Pham Nguyen Anh

Huy

Nguyen Quang Huy

DECLARATION OF AUTHORSHIP

We hereby declare that this is our own project and is guided by Dr. Hoang Anh. The content research and results contained herein are central and have not been published in any form before. The data in the tables for analysis, comments and evaluation are collected by the main author from different sources, which are clearly stated in the reference section.

In addition, the project also uses some comments, assessments as well as data of other authors, other organizations with citations and annotated sources.

If something wrong happens, we'll take full responsibility for the content of my project. Ton Duc Thang University is not related to the infringing rights, the copyrights that We give during the implementation process (if any).

Ho Chi Minh city, 23th December 2025.

Author

(Signature and full name)

Nhat

Hoang Le Minh Nhat

Anh

Pham Nguyen Anh

Huy

Nguyen Quang Huy

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1. INTRODUCTION

1.1 Problem Overview

With the rapid growth of e-commerce in Vietnam, smartphone reviews have become a vital resource for consumers. However, Traditional Sentiment Analysis (TSA), which classifies an entire document as positive or negative, often fails to capture the necessary nuance of user feedback. Specifically, when customers review different aspects or features of a phone, a document-level or sentence-level sentiment score does not effectively serve the purpose, for instance, when a user might praise the Camera but complain about the Battery. This limitation necessitates the development of Aspect-Based Sentiment Analysis (ABSA).

Figure 1.1.1 illustrates a typical product review section on an e-commerce platform, and the crucial distinction between the output of TSA and ABSA for a nuanced review is detailed in Table 1.1.1.

Input	**TSA Output**	**ABSA Output**
Điện thoại này có pin khá trâu nhưng màn hình xấu	NEUTRAL	BATTERY#POSITIVE SCREEN#NEGATIVE
This phone has a pretty long-lasting battery but bad at screen		

Table 1.1.1. Traditional Sentiment Analysis and Aspect-Based Sentiment Analysis Comparison

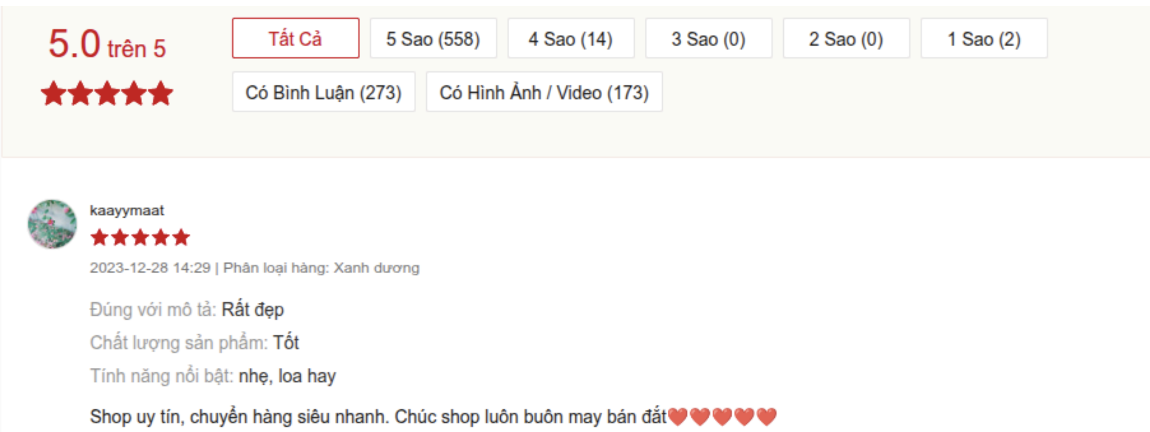

Figure 1.1.1. Illustrative image of the product review section (Shopee.vn)

Furthermore, the Vietnamese language presents unique challenges for ABSA due to its complex word segmentation (compound words), lack of clear delimiters, and the prevalence of "teencode" (internet slang) and misspellings in social media data.

1.2 Objectives

The primary objective of this project is to develop a robust ABSA system for Vietnamese smartphone reviews. Figure 1.2.1 provides an illustrative example of the desired structured output from the ABSA system. Specifically, we aim to:

1. Implement a deep learning baseline using XLM-Roberta for contextual embedding.
2. Enhance the model with Character-level CNNs (CharCNN) and Syllable embeddings to handle out-of-vocabulary (OOV) words and Vietnamese morphology.
3. Utilize a BiLSTM-CRF decoder for sequence labeling (Span Detection and Sentiment Classification).
4. Evaluate the model on the UIT-ViSD4SA dataset.

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

nearly 3 to 4 hours to full$_{\text{BATTERY\#POSITIVE}}$. Only use to surf the web, facebook, youtube. Listening to music is very good, especially listening with AKG headphones. Well worth the money spent$_{\text{GENERAL\#POSITIVE}}$	

Figure 1.2.1. Examples illustrating aspect-based sentiment analysis in Vietnamese (KTT Nguyen)

1.3 Report structure

This report is organized as follows:

Chapter 2. Background and Related Work: Discusses the dataset and foundation models.

Chapter 3. Proposed Methodology: Details the architecture.

Chapter 4. Experiments: Describes the setup and metrics.

Chapter 5. Results and Discussion: Analyzes the performance.

Chapter 6. Conclusion: Summarizes findings and future directions.

CHAPTER 2. BACKGROUND AND RELATED WORK

2.1 UIT-ViSD4SA Dataset

We utilize the UIT-ViSD4SA dataset, a benchmark specifically curated for Vietnamese Aspect-Based Sentiment Analysis. The dataset consists of smartphone reviews annotated with span boundaries and three sentiment label tags for each identified aspect (Positive, Negative, and Neutral). The raw data is processed into the standard IOB (Inside-Outside-Beginning) format (e.g., B-CAMERA#POSITIVE) for sequence labeling tasks.

Table 2.1.1 provides the full list of ten predefined aspects relevant to smartphone reviews and their brief definitions, as introduced by the dataset's creators (Phan et al., 2021).

Aspect	**Definition**
SCREEN	The users comment about the screen quality, size, colors, or display technology

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

Table 2.1.1. The full list of ten aspects and their brief definitions (Phan et al., 2021)

2.2 Exploratory Data Analysis

To better understand the inherent challenges of the UIT-ViSD4SA dataset, we performed an exploratory data analysis focusing on class distribution and sample structure.

The dataset exhibits a significant class imbalance, as illustrated in the distribution of aspect categories and sentiment polarities in Figure 2.2.1 and Figure 2.2.2. Aspects such as "General", "Performance", and "Battery" are heavily represented, while technical aspects like "Storage" and "Screen" constitute the minority classes. Furthermore, the sentiment distribution is typically skewed towards positive reviews, a common characteristic in e-commerce data that poses a challenge for accurately identifying negative feedback. This imbalance necessitates the use of the WeightedRandomSampler method in our training pipeline to mitigate bias towards the majority classes.

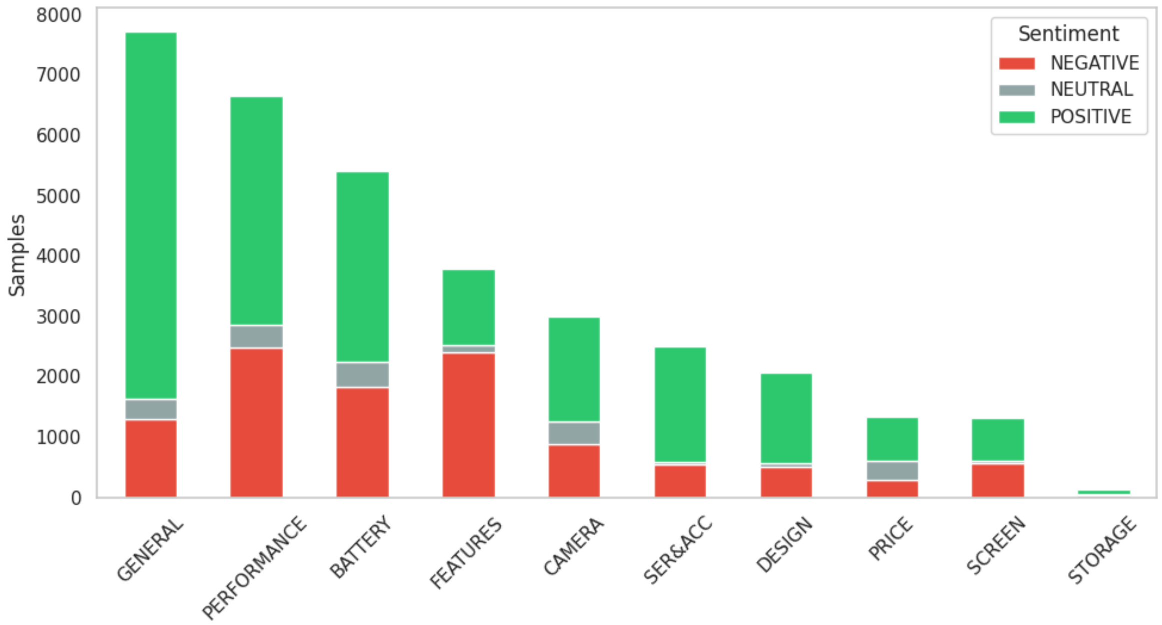

Figure 2.2.1. Distribution of Aspect and Sentiment labels in the training set

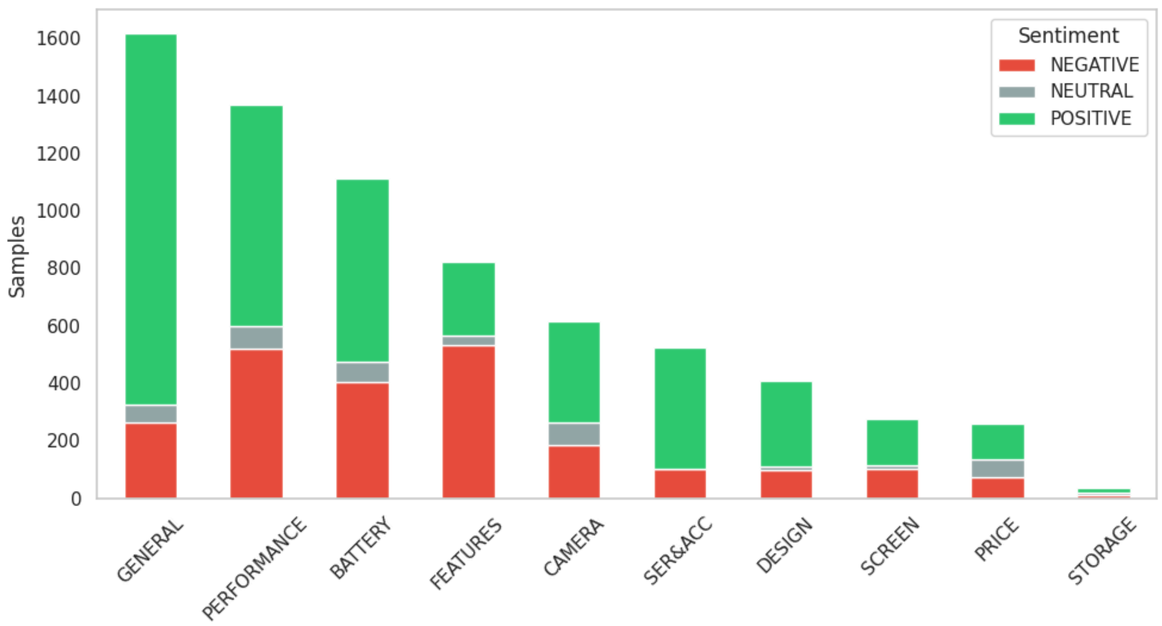

Figure 2.2.2. Distribution of Aspect and Sentiment labels in the testing set

Finally, we analyzed the review length distribution (in raw text tokens) to inform our tokenization strategy. The mean length for the training set was 36 tokens, with the 95th percentile at 78.0 tokens. Similarly, the testing set had a mean length of 36.3 tokens and a 95th percentile of 77 tokens. This distribution, illustrated in Figure 2.2.3 (Training Set) and Figure 2.2.4 (Testing Set), confirms that the majority of reviews are concise. This statistical evidence supports setting the maximum sequence length to 256 (as specified in Chapter 4) as it is sufficient to cover nearly all reviews in the dataset.

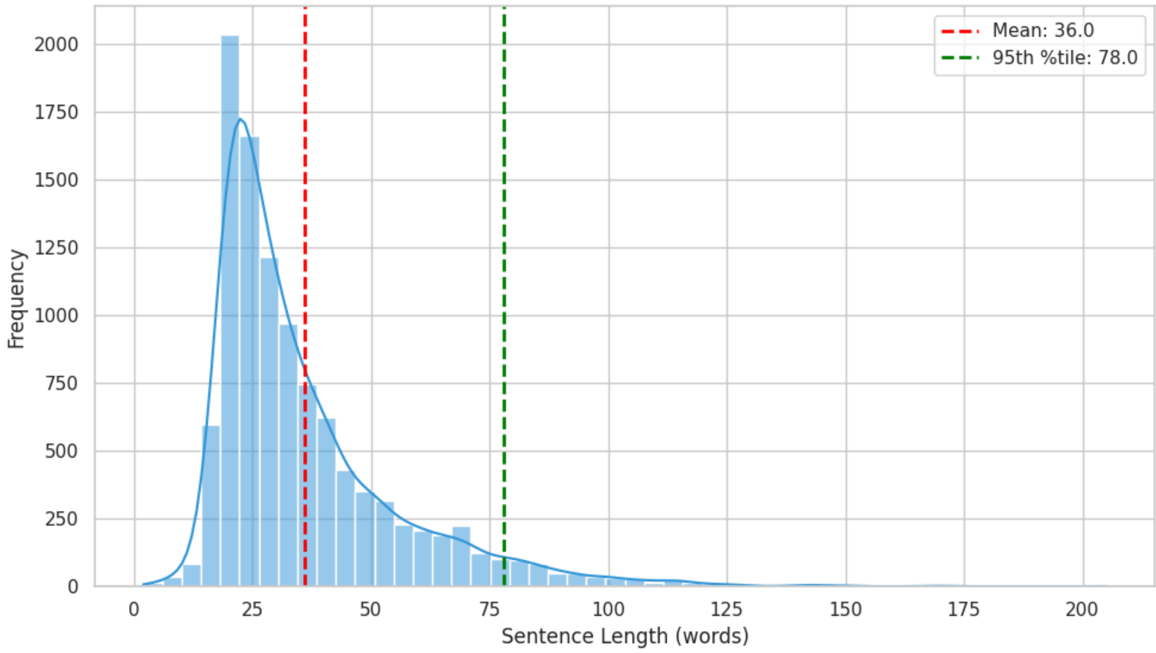

Figure 2.2.3. Reviews Length Distribution of Training Set

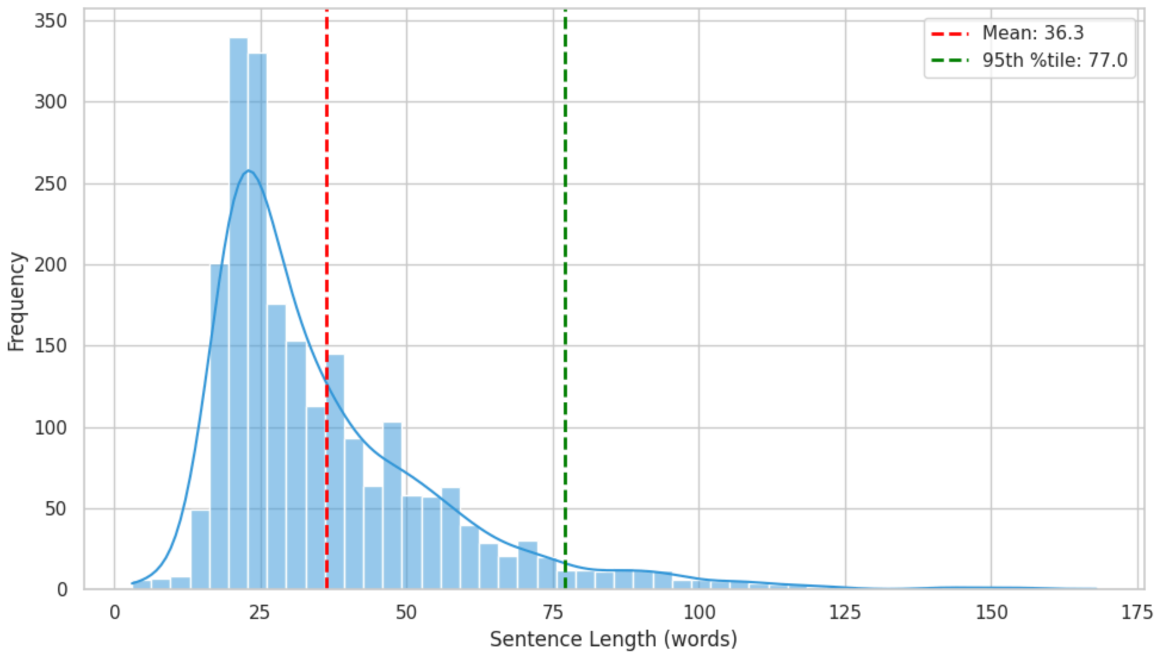

Figure 2.2.4. Reviews Length Distribution of Testing Set

2.3 Foundation Methods

Recent advancements in Natural Language Processing (NLP) rely heavily on Transformer-based models. XLM-Roberta (XLM-R) stands out as a robust multilingual model pre-trained on massive datasets, demonstrating strong performance on low-resource languages such as Vietnamese. However, standard Transformer architectures typically operate on subwords via tokenizers like SentencePiece. This can sometimes hinder the capture of fine-grained morphological details and internal word structures, which are particularly important for accurate span detection in morphologically rich or agglutinative languages like Vietnamese. This limitation motivates our proposed hybrid architecture.

2.4 Related Work

Span Detection for Aspect-Based Sentiment Analysis in Vietnamese

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

Abstract

Aspect-based sentiment analysis is a challenging task that plays an essential role in natural language processing (NLP) and artificial intelligence. However, recent works only focused on aspect detection and sentiment classification but ignored the sub-task of detecting user opinion, which has im-

more people find advice from websites, e-commerce sites, forums, or product review channels. Therefore, the number of reviews is increasing and becoming a valuable resource for customers and business. For customers, this data source provides information about products and helpful advice to help them avoid buying products or signing up for services that are not suitable for their personal needs. On the other

Figure 2.4.1. Span Detection for Aspect-Based Sentiment Analysis in Vietnamese

Our work builds upon the architectural concepts established in contemporary sequence labeling research, particularly those published in forums like ACL and PACLIC. These studies frequently suggest hybridizing strong contextual representations (from Transformer embeddings) with sequence modeling layers (like BiLSTM) and probabilistic graphical models (like Conditional Random Fields, CRF) to enforce valid and globally optimal label transitions for tasks such as named entity recognition and Aspect-Based Sentiment Analysis.

CHAPTER 3. PROPOSED METHODOLOGY

We propose a hybrid architecture that fuses semantic context with morphological features. The pipeline consists of three main stages: (1) Input Representation, (2) Context & Feature Encoding, and (3) Sequence Labeling Decoder.

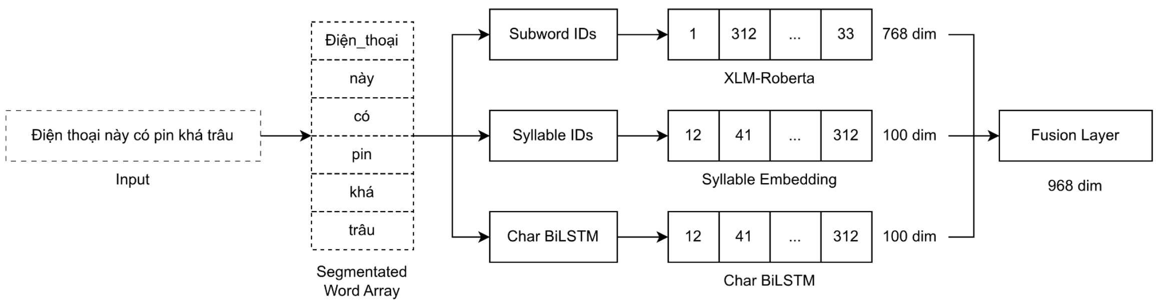

Figure 3.1. Overall Architecture

3.1 Input Representation

Given a review sentence ***S***, we first preprocess it using VnCoreNLP for word segmentation to handle Vietnamese compound words (e.g., "điện_thoại" is treated as a single semantic unit). The system then constructs three parallel input representations:

1. Subword Tokens (x_sub): Generated by the XLM-Roberta tokenizer. Since one word may be split into multiple subwords, we employ a Mean Pooling Alignment strategy. This aggregates subword embeddings back to the original word-level indices ensuring alignment with the labels.
2. Word/Syllable Indices (x_word): The segmented tokens from VnCoreNLP are mapped to a learned embedding matrix. This captures the

domain-specific vocabulary that might be fragmented by the standard subword tokenizer.

3. Character Indices (x_char): Each word is decomposed into a sequence of characters to feed into the Character-level encoder.

3.2 Feature Encoding

The core of our model fuses semantic context with morphological signals:

- Context Encoder (XLM-Roberta): We utilize xlm-roberta-base to extract contextual features (d_contex}=768). To adapt the model to the smartphone domain while preserving general linguistic knowledge, we freeze the initial layers and only fine-tune the last 4 Transformer encoder layers.
- Character-Level CNN (CharCNN-LSTM): To handle Out-of-Vocabulary (OOV) terms and misspellings (common in "teencode", e.g., "iporn" instead of "iphone"), we employ a Character-level BiLSTM. It processes character embeddings (trd_char=100) and outputs a dense vector representing the morphological structure of each word.
- Fusion Layer: The outputs from the Context Encoder, Word Embedding, and Char-Encoder are concatenated and projected through a Linear layer with ReLU activation:

$$h_{fused} = \mathrm{ReLU}(W_f \cdot [E_{xlm}; E_{word}; E_{char}] + b_f)$$

This results in a unified feature vector of dimension d_model.

3.3 Sequence Decoding (BiLSTM-CRF)

BiLSTM: A Bidirectional LSTM processes the fused features to capture long-range syntactic dependencies in both forward and backward directions.

CRF Layer: Finally, a Conditional Random Field (CRF) layer is employed to decode the sequence of labels y = (y_1, y_2, ..., y_n). The CRF optimizes the conditional probability P(y|X) by learning two types of potentials:

- Emission scores: The likelihood of a tag given the BiLSTM output.

- Transition scores: The likelihood of moving from tag y_i to tag y_i+1 (e.g., ensuring I-CAMERA follows B-CAMERA).

CHAPTER 4. EXPERIMENTS

4.1 Training Configuration

We implemented the model using PyTorch and evaluated it on the UIT-ViSD4SA dataset.

- Optimizer: We employed the AdamW optimizer with differential learning rates to prevent catastrophic forgetting in the pre-trained model:
    - theta_encoder = 1e-5 (for XLM-Roberta layers).
    - theta_decoder = 1e-3 (for LSTM-CRF head).
- Handling Class Imbalance: As identified in the EDA, classes like STORAGE and PRICE are underrepresented. We addressed this by utilizing a WeightedRandomSampler during training. Minority samples were assigned a weight of 5.0, increasing their sampling probability in each batch.
- Regularization: We applied Early Stopping with a patience of 5 epochs to prevent overfitting.

4.2 Evaluation Metrics

We strictly follow the CoNLL-2003 evaluation standard. A prediction is considered correct (True Positive) only if it matches the ground truth in all three aspects:

1. Span Boundary: Exact match of start and end indices.
2. Aspect Category: Correct aspect classification (e.g., CAMERA).
3. Sentiment Polarity: Correct sentiment (Positive/Negative/Neutral).

CHAPTER 5. RESULTS AND DISCUSSION

5.1 Performance Analysis

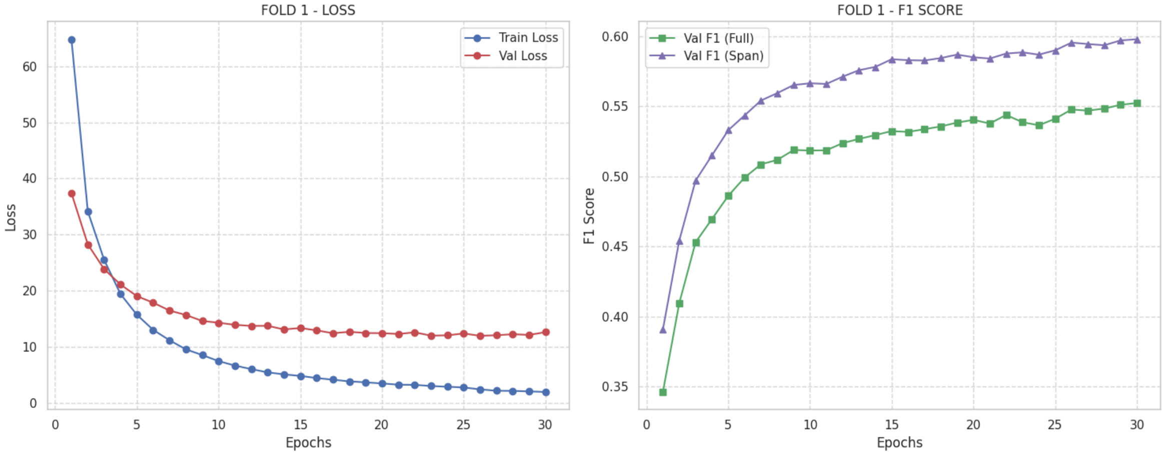

Figure 5.1.1. Training Results

The training dynamics, visualized in Figure 5.1.1, illustrate the model's convergence and generalization ability over the training epochs. The curves typically depict the loss function (Negative Log Likelihood for the CRF layer) and the Macro-Averaged F1-score for both the training and validation sets.

- Training Loss & Convergence: The training loss curve shows a steady and continuous decrease, indicating that the AdamW optimizer is successfully minimizing the objective function and the model is learning from the data. The validation loss curve closely tracks the training loss initially, suggesting good generalization.
- Validation Performance & Early Stopping: We observe that the Macro-Averaged F1-score on the validation set initially increases rapidly and then begins to plateau or show minor fluctuations. As specified in Section 4.1 (Training Configuration), the Early Stopping mechanism was successfully triggered after the validation performance ceased to improve for 5 consecutive epochs. This prevents the model from overfitting to the complexities of the training data, securing the best possible generalization score on the unseen validation data.
- Generalization: The consistent performance between the training and validation curves, up to the stopping point, confirms that the combined architecture (XLM-R + Syllable + Char-Level BiLSTM + CRF) effectively

mitigates catastrophic forgetting and class imbalance issues, leading to a stable and generalizable model. The final reported F1-score of 57.66% is the result achieved at this optimal checkpoint.

The model achieved a macro-averaged **F1-score of 57.66%** on the test set.

Aspect	Precision (%)	Recall (%)	F1-Score (%)
SCREEN	56.85	60.14	58.45
CAMERA	61.60	61.10	61.35
FEATURES	53.03	49.94	51.44
BATTERY	60.69	60.14	60.41
PERFORMANCE	55.36	55.48	55.42
STORAGE	41.38	35.29	38.10
DESIGN	63.80	61.46	62.61
PRICE	44.23	44.40	44.32
GENERAL	59.86	60.20	60.03
SER&ACC	60.48	57.82	59.12
MACRO AVERAGE	-	-	55.12

Table 5.1.1. Detailed Performance by Aspect (With CharCNN)

Best Performing Classes: DESIGN (62.61%) and CAMERA (61.35%) achieved the highest scores. These aspects often appear with distinct, descriptive adjectives (e.g., "sang chảnh", "nét", "mờ") that are effectively captured by the attention mechanism of XLM-R.

Challenges with Minority Classes: STORAGE (38.10%) and PRICE (44.32%) remain the most challenging. despite the weighted sampling. These entities often involve numerical values (e.g., "128GB", "5 triệu") which are context-dependent and ambiguous without strong surrounding cues.

Level	Precision (%)	Recall (%)	F1-Score (%)
Span Detection	61.76	61.08	61.42
Overall ABSA	57.98	57.35	57.66

Table 5.1.2. Overall with Span Detection Performance (With CharCNN)

5.2 Ablation Study

To validate the effectiveness of the Char-Encoder, we compared the full model against a version without character features.

Aspect	Precision (%)	Recall (%)	F1-Score (%)
SCREEN	58.98	54.71	56.77
CAMERA	61.73	58.83	60.25
FEATURES	50.06	46.63	48.82
BATTERY	59.59	57.45	58.50
PERFORMANCE	55.25	50.80	52.93
STORAGE	37.50	35.29	36.36
DESIGN	63.14	59.76	61.40
PRICE	45.45	38.61	41.75
GENERAL	61.32	57.23	59.21
SER&ACC	57.99	56.11	57.03
MACRO AVERAGE	-	-	53.30

Table 5.2.1. Detailed Performance by Aspect (Without CharCNN)

With Char-Encoder: F1 = 57.66%

Without Char-Encoder: F1 = 55.90% The +1.76% improvement confirms that character-level information is crucial for Vietnamese social media text, helping the model generalize better on informal spellings and compound word structures.

Level	Precision	Recall	F1-Score
Span Detection	61.79	58.04	59.86
Overall ABSA	57.71	54.21	55.90

Table 5.2.2. Overall with Span Detection Performance (Without CharCNN)

5.3 Error Analysis and Confusion Matrices

To understand the specific limitations of our model, we visualized the prediction performance using Confusion Matrices for both Sentiment (Figure 5.3.1) and Aspect (Figure 5.3.2) levels.

5.3.1 Sentiment Level Analysis

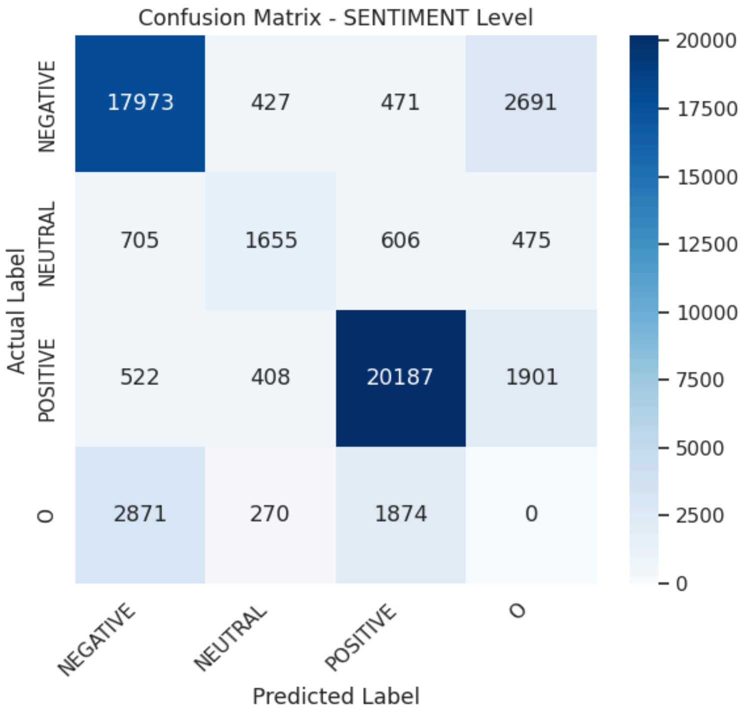

Figure 5.3.1.1. Confusion Matrix (Sentiment Level)

The Sentiment Confusion Matrix reveals significant insights regarding polarity classification challenges:

- Class Imbalance Bias: The model performs exceptionally well on the POSITIVE class (20,187 correct predictions). This is expected given the skew in the dataset towards positive reviews.
- The "Neutral" Ambiguity: The NEUTRAL class proves to be the most difficult to classify.
  - A significant number of actual NEUTRAL samples were misclassified as NEGATIVE (705 samples) and POSITIVE (606 samples).
  - Reasoning: In Vietnamese reviews, words like "ổn" (fine/okay) or "tạm được" (acceptable) are context-dependent. They can imply a positive sentiment in a budget phone review or a neutral/negative sentiment in a flagship phone review. The model struggles to capture this subtle nuance.
- Span Detection Failures (The "O" Class): There is a high rate of False Negatives where the model fails to detect sentiment entirely.

- Notably, 2,691 NEGATIVE spans and 1,901 POSITIVE spans were predicted as O (Outside). This suggests that while the model is good at classifying detected spans, it frequently misses the boundaries of sentiment-bearing phrases, particularly for negative feedback which might be expressed more implicitly (sarcasm or indirect complaints).

5.3.2 Sentiment Level Analysis

Figure 5.3.2.1. Confusion Matrix (Aspect Level)

The Aspect Confusion Matrix highlights semantic overlaps between specific smartphone features:

- Semantic Overlap (Features vs. Performance):
    - There is considerable confusion where FEATURES are misclassified as PERFORMANCE (651 samples).
    - Reasoning: This is likely due to the linguistic proximity of these concepts. For example, a "fast fingerprint sensor" (Feature) contributes to the "smoothness" of the phone (Performance). The model often conflates functional components with the resulting user experience.
- General vs. Specific Aspects: The GENERAL aspect is frequently confused with PERFORMANCE (166 misclassifications). Users often use general terms like "máy ngon" (good machine) which the model might interpret as a performance indicator rather than a general summary.
- The "Hallucination" Problem (False Positives): The column O (Predicted) shows that the model often predicts aspects where there are none (False Positives). For instance, 1,276 background tokens (Actual O) were incorrectly labeled as PERFORMANCE. This indicates the model is over-sensitive to certain adjectives (e.g., "nhanh", "mượt") and assigns them to an aspect even when they appear in a non-aspect context.
- Minority Class Performance: STORAGE shows high error rates relative to its size, often being confused with PERFORMANCE (60 samples). This confirms the difficulty of detecting technical specifications that lack strong descriptive adjectives.

CHAPTER 6. CONCLUSION AND FUTURE WORK

This study presented a robust Deep Learning approach for Aspect-Based Sentiment Analysis on Vietnamese smartphone reviews. By combining the semantic power of XLM-Roberta with the morphological awareness of Char-CNN and the structural constraints of CRF, we achieved a competitive F1-score of 57.66%.

Future Work: To further enhance the system, we propose:

1. Data Augmentation: generating synthetic samples for minority classes (STORAGE, PRICE) using synonym replacement or back-translation.
2. Post-Processing Rules: Implementing heuristic rules to correct common span boundary errors based on Part-of-Speech (POS) tags.
3. Graph Neural Networks (GNN): Incorporating dependency trees to explicitly model the syntactic distance between aspect terms and sentiment words, addressing the issue of long-range dependencies.