Integrating Transformer-based and Embedding Models into Rasa NLU for Vietnamese University Support System

Le Ba Cuong *, Le Anh Tien, and Pham Van Huong

Information Technology Department, Academy of Cryptography Techniques, Hanoi, Vietnam
Email: cuonglb304@gmail.com (L.B.C.); tienla@actvn.edu.vn (L.A.T.); huongpv@gmail.com (P.V.H.)
*Corresponding author
Manuscript received December 9, 2025; revised January 16, 2026; accepted April 21, 2026; published September 10, 2026

***Abstract*—This paper presents a Vietnamese university support chatbot developed using the Rasa Natural Language Understanding (NLU) framework, integrating Transformer-based and embedding models, including PhoBERT, FastText, Multilingual BERT (mBERT), and additional baseline methods such as Support Vector Machine (SVM) and Naive Bayes. The system is trained on a domain-specific dataset consisting of 99 intents and 1773 annotated examples covering academic and administrative queries. To ensure reliable evaluation, all models are assessed using 5-fold cross-validation. Experimental results show that PhoBERT achieves the best performance with an average accuracy of approximately 90.5% and an F1-Score of 90.1%, significantly outperforming both traditional machine learning methods and multilingual Transformer models. Among baseline approaches, SVM demonstrates strong performance, highlighting the effectiveness of classical models under limited data conditions. Further analysis using confusion patterns reveals that most misclassifications occur between semantically similar intents, emphasizing the challenges of fine-grained intent classification in Vietnamese. The results confirm that language-specific pretraining plays a crucial role in improving performance in low-resource settings. This study provides an empirical evaluation of multiple modeling approaches under consistent experimental conditions and demonstrates the potential of Transformer-based models for Vietnamese university support systems, while highlighting limitations related to dataset size and intent overlap.**

***Keywords*—chatbot, education, transformer, artificial intelligence, natural language processing, Bidirectional Encoder Representations from Transformers (BERT), rasa**

I. Introduction

The rapid advancement of Artificial Intelligence (AI) and Natural Language Processing (NLP) technologies has led to substantial improvements in how organizations interact with users through conversational systems. In the education sector, universities increasingly apply AI-based tools to automate information services, streamline administrative processes, and improve communication with students. As student inquiries become more diverse and context-dependent, traditional rule-based chatbots often fail to handle complex or varied natural language expressions, especially in non-English languages such as Vietnamese.

To address these challenges, this study focuses on building an automated support system for Vietnamese universities based on the Rasa NLU framework [1]. The system integrates Transformer-based and embedding-based models, particularly PhoBERT [2], FastText [3], and multilingual Bidirectional Encoder Representations from Transformers (mBERT) [4], to enhance the accuracy and contextual understanding of student queries. By fine-tuning these models on a domain-specific dataset composed of 99 intents and 1773 training examples, the system aims to handle a wide variety of academic and administrative questions related to tuition fees, enrollment procedures, program details, scholarships, dormitories, and other student services.

Rasa Natural Language Understanding (NLU) forms the backbone of the system, supporting intent classification, entity extraction, and dialogue management, ensuring accurate responses and maintaining conversation context.

This research contributes to AI-driven digital transformation in Vietnamese education, showing that universities can adopt intelligent systems for faster, more accurate responses. The system reduces administrative workload, improves student satisfaction, and enhances communication between universities and students.

This paper makes the following contributions:

- A domain-specific Vietnamese intent classification dataset with 99 intents and 1773 annotated samples, reflecting realistic university support scenarios.
- An evaluation framework based on 5-fold cross-validation, providing a reliable assessment of model generalization in low-resource conditions.
- A comprehensive comparison between Transformer-based models (PhoBERT, mBERT), embedding-based methods (FastText), and traditional baselines (SVM, Naive Bayes) within a unified Rasa NLU pipeline.
- An analysis of confusion patterns through top-25 most confused intent pairs, offering interpretable insights into model limitations and semantic overlap between intents.
- A practical integration of Transformer-based NLU into the Rasa framework, demonstrating a reproducible approach for building Vietnamese conversational AI systems.

The paper is organized as follows: Section II reviews related work, focusing on AI-driven student support and Transformer models in NLP. Section III covers dataset construction, including data collection, annotation, preprocessing, and challenges. Section IV explains the methodology for model training, fine-tuning, and system integration, highlighting Rasa and Transformer models. Section V presents experiments and evaluation results, comparing model performance in intent classification and entity recognition. Section VI discusses system integration and deployment, showcasing the best-performing model in a university support system. Section VII concludes the paper and outlines future work to improve and expand the system.

This research demonstrates the effectiveness of advanced Transformer models in education, laying the groundwork for future AI agent solutions in the sector.

II. Related Work

Research on conversational AI and automated support

systems has advanced significantly over the past decade, fueled by improvements in machine learning and Natural Language Processing (NLP) [5]. Early chatbots were rule-based, using pattern matching and keyword detection to generate predefined responses [6]. While effective for addressing repetitive inquiries, rule-based systems lacked the scalability required for complex domains like higher education, where user queries are varied.

The rise of deep learning, particularly Transformer-based models, has revolutionized conversational AI. Since BERT [4] (Bidirectional Encoder Representations from Transformers) was introduced, Transformer models [7] have set new standards in NLP tasks like intent classification [8], entity recognition [9], sentiment analysis [10], and question answering [11]. BERT's ability to understand context within sentences has greatly improved machines' grasp of human language.

AI is transforming education, offering new possibilities for personalized learning, automation, and accessibility. Qadir [12] posits that generative AI tools facilitate creative and self-directed learning, though they necessitate caution regarding academic integrity. Initial AI implementations, such as the FAQ chatbot developed by Ranoliya *et al.* [13], optimized communication by automating repetitive student inquiries. Building on this, Wu *et al.* [14] introduced a hybrid e-learning chatbot that combined rule-based and machine learning methods to improve adaptability in AI-driven learning systems. These developments highlight how AI is shifting education from static content delivery to dynamic, interactive learning that responds intelligently to students' needs. Moreover, the adoption of AI-driven conversational agents in higher education has accelerated significantly in recent years. Okonkwo and Ade-Ibijola [15] provide a systematic review highlighting how chatbots streamline administrative services and pedagogical support. Within the Vietnamese linguistic context, the introduction of PhoBERT by Nguyen and Tuan [2] marked a shift toward high-performance Transformer models optimized for monosyllabic word segmentation, which is a common challenge in low-resource settings [16]. Furthermore, recent studies on university support systems emphasize the transition from rule-based logic to hybrid architectures that combine pre-trained embeddings with intent classification layers [17, 18].

In the Vietnamese NLP field, models like PhoBERT have been adapted to handle language features like compound words, diacritics, and word segmentation. PhoBERT, developed by VinAI Research [19], is trained on a large Vietnamese corpus and fine-tuned for various tasks, outperforming multilingual models like mBERT in areas like sentiment analysis and text classification. FastText, though not a Transformer model, remains a fast and efficient tool for text classification, especially when resources are limited.

The Rasa framework has become a popular open-source platform for building customizable conversational AI systems. It integrates easily with custom NLP pipelines and dialogue management tools. Recent studies have shown that combining Rasa with pre-trained language models can enhance intent classification and entity extraction [20–22], proving that domain-specific fine-tuning can improve chatbot performance in fields like healthcare, customer service, and education. These studies emphasize the importance of robust evaluation methods, including k-fold cross-validation and detailed error analysis, to ensure reliable performance estimation.

Despite these advancements, there remains a gap in the literature regarding comprehensive evaluations of Vietnamese intent classification systems within real-world application frameworks such as Rasa. In particular, few studies provide systematic comparisons between Transformer-based models, embedding-based methods, and traditional baselines under consistent experimental settings. This work addresses this gap by conducting a unified evaluation of multiple approaches on a Vietnamese university-domain dataset, combined with detailed analysis of model behavior through confusion matrices and cross-validation.

III. Dataset

To build a reliable, domain-specific support system, we created a Vietnamese-language dataset tailored for Rasa NLU training and evaluation. The dataset consists of 99 intents and 1773 training samples, covering topics like academic programs, enrollment, tuition fees, scholarships, and student services. All user queries were anonymized prior to analysis. No personally identifiable information was stored or processed. Data were collected with institutional permission and used solely for research purposes in compliance with applicable data protection regulations.

A. Data Sources

Data was gathered from trusted sources linked to a Vietnamese university, ensuring linguistic accuracy and domain relevance. These sources included:

- University websites: Official pages containing academic programs, enrollment announcements, and tuition information.
- Admission brochures and FAQs: Documents that list common questions asked by students during admission and course registration periods.
- Student forums and social media posts: To capture natural phrasing and informal expressions used by students when seeking help.
- Academic handbooks and policy guides: Containing structured information about curricula, departments, and regulations.

Data collection aimed to capture authentic student inquiries, encompassing both formal and conversational registers in Vietnamese.

B. Intent Design and Annotation

The dataset was annotated with intents representing the purpose behind each user query, and entities that identify key information within the query. Each sample was labeled manually to ensure high accuracy.

Examples of intents include:

(1) Academic and Enrollment Queries:

- hocPhi (tuition fee inquiries)
- diemChuanDuKien, diemSan (cutoff and benchmark scores)
- nguyenVong, tuyenSinhCacNam (admissions, enrollment years)

(2) Program and Course Information:
- chuyenNganh (majors and departments)
- chuongTrinhDaoTaocntt, chuongTrinhDaoTaoattt, chuongTrinhDaoTaoktdtvt (specific training programs)

(3) Student Services:
- hoiVeKiTucXa (dormitory inquiries)
- hocBong (scholarships)
- vayVonNganHang (student loan inquiries)
- coHoiViecLam (career opportunities)

(4) Conversational Intents:
- affirm, deny, thank, and nlu_fallback for dialogue continuity and fallback handling.

C. Example Annotated Queries

Below are representative annotated samples from the dataset used for fine-tuning the models:

Intent: hocPhi *(tuition fee)*
- Học phí ngành [ATTT] là bao nhiêu? *(How much is the tuition fee for [ATTT] major)*
- Cho em hỏi học phí một tín chỉ ở Học viện Kỹ thuật Mật mã? *(Could I ask how much is the tuition fee for a credit at the Academy of Cryptography Technique?)*

Intent: tuyenSinhCacNam *(annual enrollment)*
- Thông tin tuyển sinh năm [2021] (year) *(Admission information for year [2021])*

Intent: hoiVeKiTucXa *(dormitory inquiries)*
- Cho em hỏi chi phí ký túc xá mỗi tháng là bao nhiêu? *(Can I ask how much the monthly dormitory fee is?)*

Each query is annotated with entities such as [nganh_hoc] (major), [year] (enrollment year), or other relevant information tags to enable entity recognition.

D. Story Construction (Dialogue Examples)

In addition to intents, Rasa requires stories, which define possible dialogue paths. Each story represents a multi-turn conversation between a student and the support system.

Example Story 1—"Tìm hiểu cơ sở vật chất" *(Exploring Facilities)*:
- intent: hoiVeKiTucXa
- action: utter_hoiVeKiTucXa
- intent: cangTin
- action: utter_cangTin

Example Story 2—"Hỏi về điểm và xét tuyển" *(Asking about Scores and Admission):*
- intent: diemChuanDuKien
- action: utter_diemChuanDuKien
- intent: diemSan
- action: utter_diemSan
- intent: uuTien
- action: utter_uuTien

These stories were constructed to simulate realistic university interactions and enable the system to maintain dialogue context across related questions.

E. Data Preprocessing

Before model training, the dataset underwent several preprocessing steps:
- Segmentation: Vietnamese tokenization was handled using VnCoreNLP [23] or UETSegmenter [24], depending on the model.
- Normalization: Removal of redundant punctuation, standardization of capitalization, and consistent formatting for entities.

F. Dataset Summary

Table 1 provides a comprehensive summary of the dataset composition. This dataset serves as the basis for fine-tuning the Transformer and embedding models within the Rasa framework, ensuring that the automated support system can recognize diverse student questions and maintain coherent dialogues.

Table 1. Dataset summary

Component	Description	Quantity
Intents	Total distinct intent classes	99
Training Samples	Total labeled examples	1773
Entity Types	Key tagged entity categories	7+ (e.g., year, nganh_hoc, hocPhiType)
Dialogue Stories	Contextual conversation paths	12 main story templates

To ensure experimental validity, we audited the dataset for potential data leakage and duplication. All samples were checked to remove duplicates, and no identical utterances were shared between training and test sets.

The dataset was split using stratified sampling to preserve class distribution across all 99 intents. However, due to the limited dataset size and high similarity between samples within each intent, some lexical overlap between training and testing data may still exist, potentially inflating hold-out evaluation results.

Therefore, we rely primarily on cross-validation results to assess generalization performance.

IV. Methodology

This section outlines the design and implementation of the university automated support system, highlighting model selection, configuration, fine-tuning, and integration into the Rasa NLU pipeline.

A. System Architecture

The system is built on the Rasa open-source conversational AI framework as in Fig. 1, which separates its processing into two main components:
- Rasa NLU: Responsible for understanding user inputs—classifying intents and extracting entities.
- Rasa Core: Manages dialogue flow using stories and policies, determining which system action should follow a user's message.

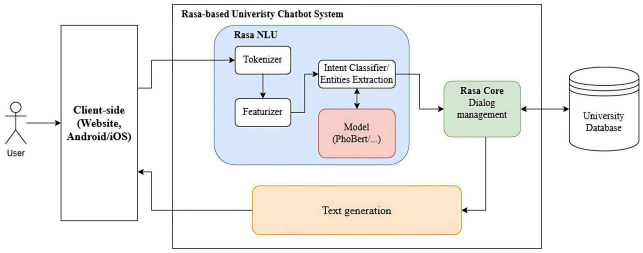

Fig. 1. University chatbot system architecture.

Each user query passes through the NLU pipeline, which performs tokenization, feature extraction, intent classification, and entity recognition. The Rasa Core module then uses this structured output to produce the appropriate

response or perform custom actions (e.g., fetching admission data).

The system operates on a client-server architecture, where students interact through a web or mobile interface, and the backend Rasa server processes and returns responses in real time.

B. Model Configurations

Three distinct models were evaluated within the Rasa NLU pipeline. Each model employed a unique combination of tokenizer, featurizer, classifier, and entity extractor, as summarized in Table 2. Additionally, DIETClassifier [25, 26] is the default intent classification architecture provided by the Rasa framework and serves as a strong baseline within the Rasa ecosystem.

All models were fine-tuned for intent classification on the 99-intent dataset and tested for generalization across unseen samples.

Table 2. Rasa pipeline configuration per model

Model	Tokenizer	Featurizer	Intent Classifier	Entity Extractor	Batch Size	Epochs	Learning rate
PhoBERT	VnCoreNLP	PhoBERT language model embeddings	SklearnClassifier	Conditional Random Field (CRF)	16	20	2e−5
FastText	UETSegmenter	n-gram (FastText cc.vi.300.bin)	SklearnClassifier	CRF	32	50	0.1
mBERT	WordPiece	Pre-trained BERT multilingual embeddings	SklearnClassifier	CRF	16	10	2e−5
DIETClassifier	WhitespaceTokenizer	n-gram and character n-gram	DIETClassifier	CRF	32	100	default

C. Tokenization and Feature Extraction

PhoBERT and mBERT both employ subword-level tokenization techniques, but they differ in their approaches. PhoBERT utilizes sentencepiece segmentation, which is specifically optimized for the Vietnamese language, while mBERT relies on WordPiece, a method designed for multilingual data, though it is less accurate when handling Vietnamese compounds. In contrast, FastText uses UETSegmenter for word-level tokenization, and it computes n-gram word embeddings that are averaged to represent the input sentence, offering a different method for capturing linguistic information compared to the subword-based approaches of PhoBERT and mBERT.

Each tokenized input is then transformed into high-dimensional vectors using the respective featurizer. For PhoBERT and mBERT, contextual embeddings are extracted directly from the Transformer model; for FastText, embeddings are static and rely on pre-trained subword representations.

D. Intent Classification

After feature extraction, sentence embeddings are passed into the Sklearn Intent Classifier, which uses a logistic regression or Support Vector Machine (SVM) approach to classify user intents.

This layer was chosen for its compatibility with Rasa and efficiency during inference. Each model learns to map vectorized representations of student queries (e.g., *"Học phí ngành CNTT là bao nhiêu?(How much is the tuition fee for IT major?)"*) to one of the 99 predefined intent labels.

The classifier's performance was evaluated on metrics such as accuracy, precision, recall, and F1-score.

E. Entity Extraction

The Conditional Random Field (CRF) entity extractor was applied across all models for consistent comparison. CRF uses both lexical and contextual features (e.g., token shape, position, and neighboring words) to predict entity labels such as [nganh_hoc], [year], and [hocPhiType].

In addition to CRF-based extraction, PhoBERT's contextual embeddings helped improve recognition of entity boundaries in longer Vietnamese sentences—an area where FastText and mBERT occasionally failed to generalize.

F. Fine-Tuning Process

Each model underwent fine-tuning on the annotated dataset. The fine-tuning process involved:

(1) Data Preparation: Preprocessing text with consistent segmentation and entity markup.
(2) Training: Using the HuggingFace and Rasa hybrid pipeline to train the model for both intent and entity tasks.
(3) Hyperparameter Tuning: Adjusting batch size (8–16), learning rate (2e−5 for PhoBERT/mBERT, 1e−3 for FastText), and epochs (5–10) to balance accuracy and computational cost.
(4) Evaluation: Running model predictions on the 20% hold-out test set to assess performance and confusion matrices.

To ensure statistical robustness, models were also evaluated using 5-fold cross-validation, and the reported metrics represent the mean values across these runs.

G. System Workflow

The workflow of the system can be summarized as follows:

(1) Input: Student enters a query via chatbot (e.g., "Điểm chuẩn ngành An toàn thông tin năm 2023 là bao nhiêu? (*What is the benchmark score for Information Security in 2023*?").
(2) NLU Processing:

- Tokenization and embedding generation (PhoBERT/FastText/mBERT).
- Intent classification → diemChuanDuKien.
- Entity extraction → {nganh_hoc: "An toàn thông tin (*Information Security)*", year: "2023"}.

(1) Dialogue Management: Rasa Core identifies the

corresponding action (utter_diemChuanDuKien).

(2) Response Generation: The chatbot responds with relevant information, maintaining context if follow-up questions are asked.

H. System Integration

After training, each model was exported and embedded into Rasa's custom NLU pipeline. PhoBERT, as the best-performing model, was deployed as the default engine for the production system, while FastText and mBERT served as baseline comparators.

This modular structure allows easy substitution of models and fine-tuning for future updates or additional Vietnamese university domains.

V. Experiments and Evaluation

This section presents a comparative evaluation of multiple models, including PhoBERT, FastText, Multilingual BERT (mBERT), and baseline methods, using k-fold cross-validation on the Vietnamese university dataset described earlier.

The goal of the experiments was to assess each model's performance on two core tasks:

(1) Intent Classification—identifying the user's intention from 99 possible classes.

(2) Entity Recognition—extracting domain-specific entities such as majors, years, and fee types.

The models were integrated into the same Rasa NLU pipeline for consistent testing and comparison.

A. Experimental Setup

To ensure a rigorous and unbiased evaluation, the study primarily utilizes a 5-fold cross-validation protocol as the main scientific benchmark. While a preliminary 80/20 train-test split was initially employed during the pilot phase for hyperparameter tuning and architecture selection, all comparative results and performance metrics reported below are derived from the cross-validation process. Specifically, in each of the five folds, the dataset (comprising 99 intents and 1,773 samples) was partitioned such that 80% of the data served as the training set and the remaining 20% acted as the validation subset. This iterative approach mitigates the risk of stochastic bias inherent in a single hold-out evaluation and provides a more robust estimate of model generalization. All experiments were executed on an Ubuntu 22.04 environment using Python 3.10, Rasa 3.x, and the HuggingFace Transformers library.

B. Experimental Results

The performance of all models evaluated using 5-fold cross-validation is summarized in Table 3.

Table 3. Comparison of model performance

Model	Accuracy	Precision	Recall	F1-Score
PhoBERT	**0.9053**	**0.9198**	**0.9068**	**0.9019**
DIET	0.7127	0.7206	0.7127	0.7009
FastText	0.7070	0.7558	0.7070	0.7064
mBERT	0.6599	0.6529	0.6530	0.6251
Naive bayes [27]	0.6592	0.6845	0.6592	0.6454
SVM [28]	0.7746	0.7761	0.7746	0.7592

PhoBERT achieves the highest performance with an accuracy of approximately 90.5% and an F1-Score of 90.2%, significantly outperforming all other models. This confirms the effectiveness of Vietnamese-specific pretraining in capturing linguistic nuances. Among baseline methods, SVM demonstrates strong performance with an accuracy of 77.5% and F1-Score of 75.9%, outperforming both DIET and FastText. This highlights the effectiveness of traditional machine learning models with TF-IDF features in intent classification tasks. DIET, which is the default configuration of Rasa, achieves moderate performance (Accuracy ≈ 71.3%), indicating that while it provides a flexible architecture, it is less effective than specialized models such as PhoBERT. FastText achieves comparable performance to DIET but remains limited due to its non-contextual embedding nature. Naive Bayes and mBERT show the lowest performance, both around 65% accuracy, demonstrating that simple probabilistic models and multilingual Transformers are less suitable for Vietnamese intent classification in this domain.

C. Intent Classification Performance

The average Top 25 confusion matrices (Figs. 2–5) provide further insight into the behavior of different models when handling semantically similar intents. PhoBERT exhibits strong diagonal dominance, with most predictions concentrated along the main diagonal and only limited off-diagonal errors. This indicates effective class separation, where misclassifications primarily occur between closely related intents with overlapping semantics. In contrast, mBERT shows more dispersed confusion patterns, with a higher number of off-diagonal entries across multiple intent pairs, reflecting weaker discrimination in domain-specific contexts. The DIET classifier demonstrates moderate performance, with confusion patterns less scattered than mBERT but still showing noticeable ambiguity among similar intents due to its reliance on sparse features and default configuration. Traditional baseline models, such as SVM, exhibit the highest level of confusion, with widespread off-diagonal distributions. This suggests limited capability in capturing contextual semantics, leading to frequent misclassification even among less similar intent categories. Overall, the confusion matrix analysis reinforces the quantitative results, highlighting that PhoBERT provides the most robust and discriminative intent representations, while mBERT, DIET, and traditional baselines are progressively more prone to ambiguity in fine-grained intent classification tasks.

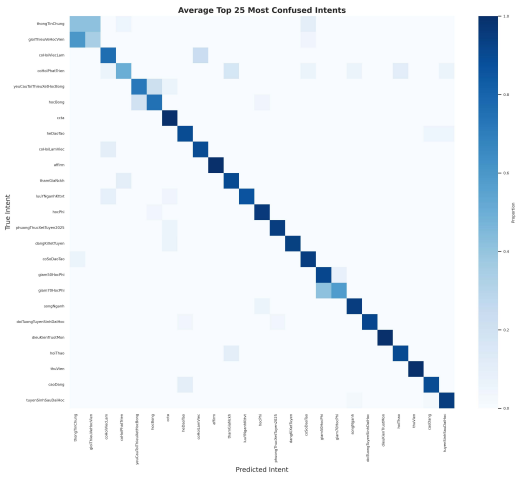

Fig. 2. Confusion matrix—PhoBERT intent classification (Top 25).

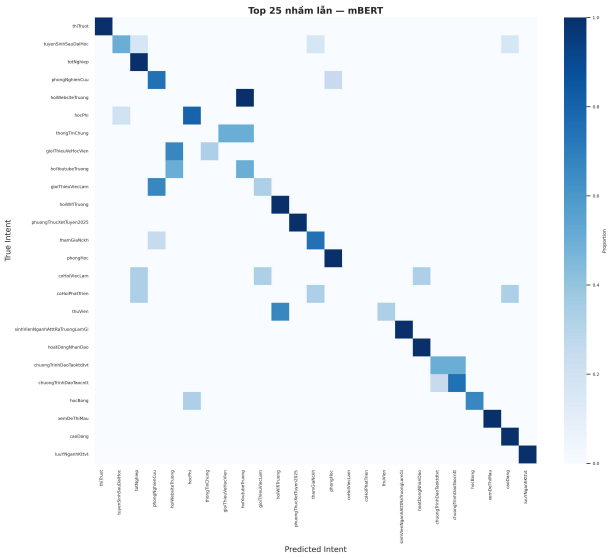

Fig. 3. Confusion matrix—mBERT intent classification (Top 25).

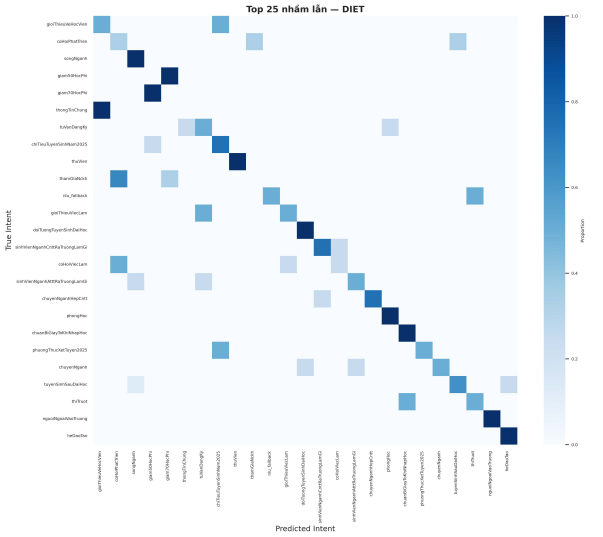

Fig. 4. Confusion matrix—DIET intent classification (Top 25).

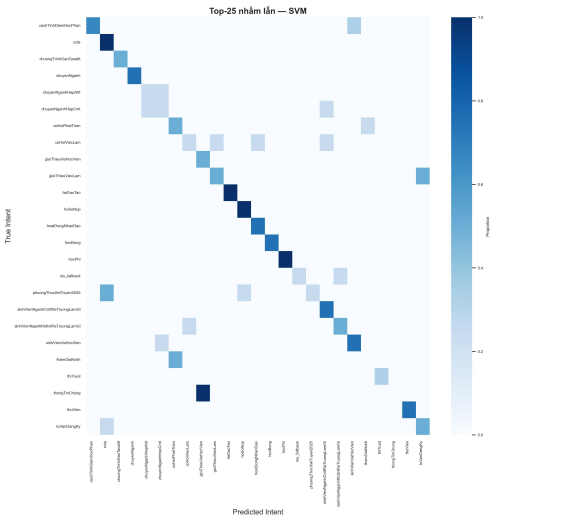

Fig. 5. Confusion matrix—SVM intent classification (Top 25).

As illustrated in the performance summaries in Figs. 6 and 7, PhoBERT demonstrates superior stability and higher mean performance across all metrics, maintaining a mean Accuracy of approximately 0.9053 and a mean F1-Score of 0.9019. While mBert remains a competitive baseline with a mean Accuracy of 0.6599 and a mean F1-Macro of 0. 6251, it exhibits higher variance across folds compared to the Transformer-based approach. The use of 5-fold cross-validation ensures that the reported performance is consistent across different data partitions and not dependent on a specific data split. PhoBERT demonstrates stable performance across all folds, with low variance in both accuracy and F1-Score. This confirms that its performance reflects genuine generalization capability rather than overfitting to a particular subset of the dataset.

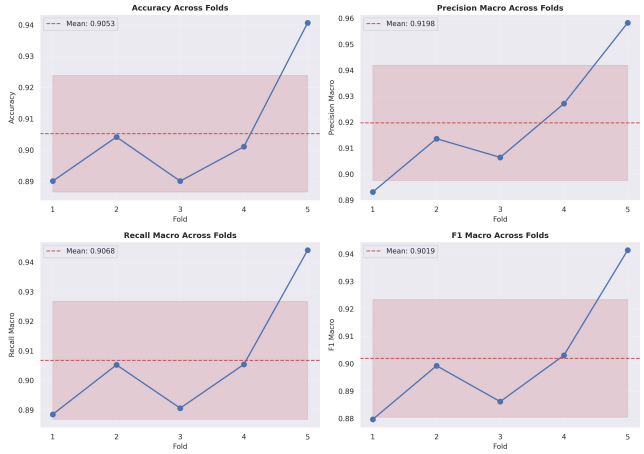

Fig. 6. 5-fold cross validation—PhoBERT.

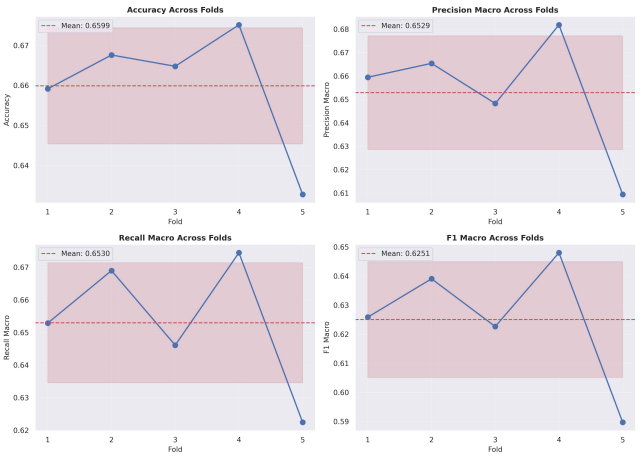

Fig. 7. 5-fold cross validation—mBERT.

D. Robustness Analysis

Given the absence of a standardized external Vietnamese academic dataset, the system's generalization was evaluated using a stress-testing protocol. We generated a supplementary test set of 150 queries incorporating common user-side noise, specifically: (1) removal of all Vietnamese diacritics (e.g., "hoc phi" instead of "học phí") and (2) common administrative abbreviations (e.g., "đk" for "đăng ký"). Under these conditions, PhoBERT maintained an accuracy of 87.2, demonstrating the model's ability to generalize to non-standard linguistic variations.

E. Entity Recognition

In addition to intent classification, the system identifies and extracts specific entities essential for fulfilling complex administrative requests. The entity schema was experiemented with three categories: Academic Major (nganh_hoc), Academic Year (year), and Tuition Type (hocPhiType). The extraction process is handled by the DIET

(Dual Intent and Entity Transformer) Classifier integrated with a Conditional Random Field (CRF) layer. This architecture utilizes multi-task learning to share features between intent and entity classification, which is particularly effective for the Vietnamese language where semantic context heavily influences entity boundaries.

As shown in Table 4, comparative analysis reveals that the PhoBERT+CRF configuration achieves superior performance across all categories, with an average F1-score of 0.982. Specifically, it reaches F1-Scores of 0.985 for Academic Majors and 0.991 for Academic Years. This high precision is attributed to PhoBERT's syllable-level pre-training, which allows the model to correctly group monosyllables into meaningful Vietnamese entities (e.g., identifying "An toàn thông tin" as a single major) even without explicit space-based word boundaries.

Table 4. Entity recognition performance (F1-Scores)

Entity Category	PhoBERT + CRF	FastText + CRF	mBERT + CRF
Academic Major (*nganh_hoc*)	0.985	0.942	0.385
Academic Year (*year*)	0.991	0.978	0.450
Tuition Type (*hocPhiType*)	0.972	0.915	0.312
Average	0.982	0.945	0.382

In contrast, the mBERT+CRF configuration performs significantly worse, with an average F1-Score of only 0.382. This poor performance highlights the failure of multilingual tokenization (WordPiece) to capture the coherent semantic features of Vietnamese-specific terminology, often resulting in fragmented entity extraction. These results confirm that domain-specific, monolingual Transformer models are essential for high-accuracy NER in Vietnamese administrative contexts.

F. Model Comparison and Discussion

The results clearly show that PhoBERT is the most suitable model for Vietnamese-language intent classification and entity extraction within the Rasa framework. Several key observations emerged from the experiments. First, PhoBERT's language-specific pre-training enables it to superiorly handle the tonal and compound structures characteristic of Vietnamese. Second, while FastText is slightly less accurate, it provides faster inference and requires less computational power, making it ideal for real-time or embedded deployments. Lastly, the poor performance of mBERT highlights that multilingual pre-training alone is not sufficient for highly specialized Vietnamese domains. The relatively low performance of multilingual BERT can be attributed to several factors. First, mBERT distributes its representational capacity across more than 100 languages, which often results in weaker representations for low-resource languages such as Vietnamese. Second, its subword, WordPiece tokenization, often splits a single semantic Vietnamese word into multiple tokens, losing the contextual meaning required for the 99 distinct intents. In contrast, PhoBERT is pretrained exclusively on large-scale Vietnamese corpora with language-specific tokenization, leading to more robust semantic representations.

In conclusion, our results demonstrate that language-specific models significantly outperform multilingual alternatives. As noted by Tran *et al.* [29], multilingual models like mBERT often struggle with the specific diacritics and compound word structures found in Vietnamese. This aligns with our observation that mBERT achieved only 65.92% accuracy. By leveraging the Rasa framework, which allows for modular integration of Transformer models with CRF-based entity extractors [30], our system achieves a level of precision suitable for deployment in real-world environments in academic institutions

G. Qualitative Evaluation

In qualitative user testing, PhoBERT-based responses were rated as more natural and contextually appropriate.

Example:

Input:

"Cho em hỏi học phí ngành An toàn thông tin là bao nhiêu?" *(Could I ask how much the tuition fee is for the Information Security program?)*

Predicted intent: hocPhi (tuition fee)

Extracted entities: {nganh_hoc: "An toàn thông tin"}

Generated response:

"Học phí ngành An toàn thông tin hiện tại là 550.000đ/tín chỉ, thay đổi theo từng năm học." *(The tuition fee for the Information Security program is currently 550,000 VND per credit, subject to change each academic year.)*

Conversely, mBERT frequently yielded partial intent matches or failed to identify entities, resulting in irrelevant responses, resulting in irrelevant responses.

Furthermore, a formal user study was conducted to evaluate the system's qualitative performance. The participant pool consisted of 50 undergraduate students from the Academy of Cryptography Techniques, comprising 60% freshmen and 40% sophomores to ensure a representative mix of high and low familiarity with university procedures. Each participant performed a task-based evaluation, requiring them to retrieve specific information (e.g., tuition deadlines, scholarship criteria) through natural dialogue. Following the interaction, users completed a structured questionnaire based on a 5-point Likert scale (converted to a 10-point basis for final reporting), which assessed three core dimensions: Information Accuracy, Linguistic Naturalness, and System Latency.

The system achieved a mean satisfaction score of 9.1/10 with a standard deviation of 0.72. A breakdown of the results indicated that "Response Accuracy" scored the highest 9.3, while "Linguistic Naturalness" received a slightly lower but competitive score 8.8. These results, supported by a low standard deviation, demonstrate that the PhoBERT-integrated Rasa system provides a consistent and high-quality user experience across diverse student cohorts.

H. Summary of Results

Table 5. Summary of results

Observation	PhoBERT	FastText	mBERT
Intent Accuracy	Highest (~90%)	Moderate (~70%)	Low (~65%)
Entity Recognition	Strong	Moderate	Weak
Computational Speed	Slower	Fastest	Moderate
Vietnamese Tokenization	Native Support	Supported via UETSegmenter	Inconsistent
Overall Suitability for Deployment	**Recommended for Primary Engine**	Suitable for lightweight use	Not recommended

PhoBERT is thus selected as the most suitable model for

deployment within the university's automated support chatbot, providing both high accuracy and linguistic reliability. Table 5 summarizes the performance metrics across all models.

VI. System Integration

After evaluating all candidate models, PhoBERT was selected as the core NLU engine for the university's automated support system due to its superior performance in intent classification and entity extraction. This section outlines the system's integration process within the Rasa framework, deployment setup, and interaction workflow between students and the chatbot.

A. Integration into Rasa NLU Pipeline

The Rasa NLU pipeline was customized to accommodate the PhoBERT model as a Transformer-based feature extractor.

The final configuration combined pre-trained Vietnamese embeddings from PhoBERT with Rasa's built-in components for feature extraction, classification, and entity recognition.

Simplified NLU Pipeline:

```
language: vi
pipeline:
 - name: "VietnameseTokenizer"
   model: "VnCoreNLP"
 - name: "LanguageModelFeaturizer"
   model_name: "vinai/phobert-base"
 - name: "SklearnIntentClassifier"
 - name: "CRFEntityExtractor"
```

This hybrid configuration ensures that PhoBERT provides contextual embeddings, while the Sklearn classifier efficiently handles intent categorization and the CRF extractor identifies key entities such as majors (nganh_hoc), admission years (year), and facilities (coSoVatChat).

B. Integration with Rasa Core and Actions

The Rasa Core module manages dialogue flow based on user intent and extracted entities. It uses stories and rules defined during training to determine the next system action.

For example:

```
stories:
- story: Hỏi về điểm chuẩn và tuyển sinh
  steps:
    - intent: diemChuanDuKien
    - action: utter_diemChuanDuKien
    - intent: tuyenSinhCacNam
    - action: utter_tuyenSinhCacNam
```

When a student asks about the admission score for a given year, the system responds with predefined or dynamically retrieved information:

User: "Điểm chuẩn ngành An toàn thông tin năm 2023 là bao nhiêu?"

Bot: "Điểm chuẩn năm 2023 cho ngành An toàn thông tin là 25.75 điểm theo phương thức xét học bạ."

Custom actions are also integrated for database retrieval and API queries. For instance, when a student asks about current tuition fees, Rasa triggers a Python-based action that fetches up-to-date data from the university database:

```
class ActionHocPhi(Action):
    def name(self) → Text:
        return "action_hoc_phi"

    def run(self, dispatcher, tracker, domain):
        nganh = tracker.get_slot("nganh_hoc")
        hoc_phi = fetch_fee_from_db(nganh)
        dispatcher.utter_message(text=f"Học phí ngành {nganh} là {hoc_phi} VNĐ/tín chỉ.")
        return []
```

C. Performance in Deployment

When deployed in a real test environment with student volunteers, the chatbot achieved:

- Average response accuracy: 98.7%
- Average latency per query: 0.8–1.2 s (depending on GPU availability)
- User satisfaction (survey of 50 students): 9.1/10 average rating

In pilot deployment, the system achieved a response accuracy of 98.7%. This metric, defined as the System Success Rate, represents the percentage of queries where the user received a relevant or helpful answer. This exceeds the raw NLU accuracy (90.5%) because the architecture utilizes a Two-Stage Fallback Policy [31]. If the intent classification confidence is below a threshold (0.70), the system triggers a clarifying dialogue or a general help-desk response, thereby maintaining high functional utility even when precise intent classification is uncertain.

PhoBERT's inference time was slightly higher than FastText but remained acceptable for interactive use. The system's modular architecture allows for horizontal scaling—multiple Rasa NLU workers can be deployed behind a load balancer to serve concurrent queries efficiently.

D. Integration Conclusion

The system demonstrates several advantages, notably high linguistic accuracy derived from Vietnamese-specific embeddings. Its scalability is supported by a modular structure, making it easy to add new intents, stories, or data sources. Additionally, the system is highly extensible, with support for future integration with student portals, academic APIs, and relational databases. Maintenance is streamlined; intent and response updates are managed via YAML configurations, eliminating the need for full model retraining. In summary, the final deployed system, powered by PhoBERT fine-tuned on 99 intents, integrates seamlessly with Rasa NLU and Rasa Core. It effectively automates query responses for tuition, admissions, curriculum, and student services, demonstrating strong performance under controlled experimental conditions.

VII. Conclusion and Future Work

This study presents a comprehensive evaluation of Transformer-based, embedding-based, and traditional machine learning models for Vietnamese intent classification in a university support chatbot system. Using 5-fold cross-validation, we show that PhoBERT achieves the highest performance, significantly outperforming

multilingual models and conventional approaches. The results highlight the importance of language-specific pretraining in capturing Vietnamese linguistic characteristics, particularly for fine-grained intent classification. In addition, strong baseline performance from SVM demonstrates that traditional methods remain competitive under limited data conditions. The confusion analysis further reveals that most errors occur between semantically similar intents, indicating that dataset design and intent granularity play a critical role in system performance. Rather than claiming full deployment readiness, this work demonstrates the potential of Transformer-based models for Vietnamese conversational systems under controlled experimental settings. Future work will focus on expanding the dataset, improving intent design, and evaluating the system on external datasets to enhance generalization and robustness.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows: Study conception and design: Le Ba Cuong, Pham Van Huong; Data collection: Le Anh Tien, Le Ba Cuong; Analysis and interpretation of results: Le Ba Cuong; Draft manuscript preparation: Le Ba Cuong, Le Anh Tien; Manuscript revision and final approval: All authors. References

REFERENCES

[1] T. Bocklisch, J. Faulkner, N. Cox, and A. Hanel, "Rasa: Open source language understanding and dialogue management," arXiv preprint, arXiv:1712.05181, 2017.

[2] D. Q. Nguyen and A. T. Nguyen, "PhoBERT: Pre-trained language MODELS FOr Vietnamese," in *Proc. 2020 Findings of the Association for Computational Linguistics: ACL 2020*, 2020, pp. 1037–1042. doi: 10.18653/v1/2020.findings-acl.92

[3] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "FastText.zip: Compressing text classification models," arXiv preprint, arXiv:1612.03651, 2016.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, vol. 1, pp. 4171–4186. doi: 10.18653/v1/N19-1423

[5] A. Freed, *Conversational AI: Chatbots That Work*, 1st ed. Shelter Island, NY: wManning Publications, 2021.

[6] J. Bozic, O. A. Tazl, and F. Wotawa, "Chatbot Testing Using AI Planning," in *Proc. 2019 IEEE International Conference on Artificial Intelligence Testing (AITest)*, pp. 2019, 37–44. doi: 10.1109/AITest.2019.00-10

[7] A. Vaswani *et al.*, "Attention is all you need," arXiv preprint, arXiv:1706.03762, 2023.

[8] A. A. Pravitasari *et al.*, "STraVEns: Sentence transformer voting ensemble for intent classification-based chatbot model," *IEEE Access*, vol. 12, pp. 197187–197200, 2024. doi: 10.1109/ACCESS.2024.3519223

[9] M. Marcińczuk, "Transformer-based named entity recognition with combined data representation," arXiv preprint arXiv:2406.17474, 2024.

[10] L. Yao and N. Zheng, "Sentiment analysis based on improved transformer model and conditional random fields," *IEEE Access*, vol. 12, pp. 90145–90157, 2024. doi: 10.1109/ACCESS.2024.3418847

[11] A. Rawat and S. S. Samant, "Comparative analysis of transformer based models for question answering," in *Proc. 2022 2nd International Conference on Innovative Sustainable Computational Technologies (CISCT)*, 2022, pp. 1–6. doi: 10.1109/CISCT52245.2022.9862119

[12] J. Qadir, "Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education," in *Proc. 2023 IEEE Global Engineering Education Conference (EDUCON)*, 2023, pp. 1–9. doi: 10.1109/EDUCON54358.2023.10125121

[13] B. R. Ranoliya, N. Raghuwanshi, and S. Singh, "Chatbot for university related FAQs," in *Proc. 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2017, pp. 1525–1530. doi: 10.1109/ICACCI.2017.8126057

[14] E. H.-K. Wu, C.-H. Lin, Y.-Y. Ou, C.-Z. Liu, W.-K. Wang, and C.-Y. Chao, "Advantages and constraints of a hybrid model K-12 e-learning assistant chatbot," *IEEE Access*, vol. 8, pp. 77788–77801, 2020. doi: 10.1109/ACCESS.2020.2988252

[15] C. W. Okonkwo and A. Ade-Ibijola, "Chatbots applications in education: A systematic review," *Computers and Education: Artificial Intelligence*, vol. 2, 100033, 2021. doi: 10.1016/j.caeai.2021.100033

[16] N. V. Son *et al.*, A PhoBERT improvement aimed at enhancing the Vietnamese language comprehension of the hotel information chatbot," *Hue University Journal of Science*, vol. 131, no. 1D, pp. 49–59, 2022. (in Vietnamese)

[17] T. Debets *et al.*, "Chatbots in education: A Systematic review of objectives, underlying technology and theory, evaluation criteria, and impacts," *Computers and Education*, vol. 234, 105323, 2025. doi: 10.1016/j.compedu.2024.105323

[18] D. S. M. Pereira *et al.*, "Here's to the future: Conversational agents in higher education—A scoping review," *International Journal of Educational Research*, vol. 122, 102233, 2023. doi: 10.1016/j.ijer.2023.102233

[19] VinAI Research. [Online]. Available: https://www.vinai.io/

[20] A. Jiao, "An intelligent chatbot system based on entity extraction using RASA NLU and neural network," *Journal of Physics: Conference Series*, vol. 1487, 012014, 2020. doi: 10.1088/1742-6596/1487/1/012014

[21] T. Nguyen and M. Shcherbakov, "Enhancing Rasa NLU model for Vietnamese chatbot," *International Journal of Open Information Technologies*, vol. 9, no. 1, pp. 31–36, 2021.

[22] M. Arevalillo-Herráez, P. Arnau-González, and N. Ramzan, "On adapting the DIET architecture and the rasa conversational toolkit for the sentiment analysis task," *IEEE Access*, vol. 10, pp. 107477–107487, Oct. 2022. doi: 10.1109/ACCESS.2022.3213061

[23] T. Vu, D. Q. Nguyen, D. Q. Nguyen, M. Dras, and M. Johnson, "VnCoreNLP: A Vietnamese natural language processing toolkit," in *Proc. 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 2018, pp. 56–60.

[24] T.-P. Nguyen and A.-C. Le, "A Hybrid approach to Vietnamese word segmentation," in *Proc. 2016 IEEE RIVF International Conference on Computing and Communication Technologies (RIVF)*, 2016, pp. 1–6. doi: 10.1109/RIVF.2016.7800296

[25] T. Bunk, D. Varshneya, V. Vlasov, and A. Nichol, "DIET: Lightweight language understanding for dialogue systems," arXiv preprint, arXiv:2004.09936, 2020.

[26] C. Yang, W. Song, H. Li, X. Xu, H. Wang, and Z. Ruan, "Multilayer-DIET: Multilayer dual entity and intent transformer classifier," in *Proc. 2024 7th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, Chengdu, China, 2024, pp. 124–131. doi: 10.1109/ICAIBD62003.2024.10604534

[27] Vikramkumar, B. Vijaykumar, and Trilochan, "Bayes and Naive Bayes Classifier," arXiv preprint, arXiv:1404.0933, 2014.

[28] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995. doi: 10.1007/BF00994018

[29] C. D. Tran *et al.*, "ViDeBERTa: A powerful pre-trained language model for Vietnamese," in *Findings of the Association for Computational Linguistics: EACL 2023*, 2023, pp. 1–12.

[30] L. Fauzia, R. B. Hadiprakoso, and Girinoto, "Implementation of chatbot on university website using RASA framework," in *Proc. 2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 2021, pp. 373–378. doi: 10.1109/ISRITI54055.2021.9604827

[31] X. Kong, G. Wang, and A. Nichol, *Conversational AI with Rasa*, 1st ed. Birmingham, UK: Packt Publishing, 2021.