XÂY DỰNG MÔ HÌNH AI CHATBOT TƯ VẤN HỌC VIÊN TẠI TRUNG TÂM NGOẠI NGỮ - TIN HỌC

Nguyễn Bá Duy[1], Trần Lê Duy Anh[1], Diệp Bình Nguyên[1], Nguyễn Thị Thúy An[2], Bùi Bích Phương[2] và Huỳnh Minh Tiến[3]

[1]Trường Đại học Kỹ thuật - Công nghệ Cần Thơ
[2]Sinh viên Khoa Công nghệ thông tin, trường Đại học Kỹ thuật - Công nghệ Cần Thơ
[3]Bảo tàng thành phố Cần Thơ
Email: nbduy@ctuet.edu.vn

Thông tin chung:
Ngày nhận bài: 29/6/2025
Ngày nhận bài sửa: 03/10/2025
Ngày duyệt đăng: 15/10/2025

Từ khóa: *Chatbot, Rasa, Tư vấn học viên.*

TÓM TẮT

Chatbot ngày nay được ứng dụng rộng rãi trong giáo dục, nhưng các mô hình ngôn ngữ lớn thường đòi hỏi hạ tầng tính toán mạnh, gây khó khăn cho các cơ sở địa phương. Nghiên cứu này giới thiệu mô hình AI Chatbot tư vấn học viên, triển khai thử nghiệm tại Trung tâm Ngoại ngữ - Tin học, Trường Đại học Kỹ thuật - Công nghệ Cần Thơ. Hệ thống sử dụng nền tảng Rasa kết hợp LaBSE và PostgreSQL, được thiết kế chạy trên CPU, không phụ thuộc GPU hay API thương mại. Bộ dữ liệu gồm 1.712 câu hỏi thực tế, phân loại thành 56 ý định và 55 kịch bản hội thoại. Kết quả đánh giá chéo (k=10) đạt độ chính xác 98,1% trong phân loại ý định và 99,9% trong nhận diện thực thể. Mô hình chứng minh hiệu quả và khả năng ứng dụng trong tư vấn học viên với hạ tầng hạn chế.

1. ĐẶT VẤN ĐỀ

Chuyển đổi số đang thúc đẩy nhu cầu ứng dụng công nghệ vào hoạt động tư vấn và hỗ trợ học viên tại các cơ sở giáo dục, đặc biệt là các trung tâm đào tạo ngắn hạn. Trong lịch sử phát triển chatbot, nhiều mô hình nền tảng đã được hình thành và ứng dụng như những chuẩn mực ban đầu để xây dựng các hệ thống về sau. Chẳng hạn, ELIZA do Weizenbaum [1] phát triển là chương trình trò chuyện dựa trên luật đầu tiên; tiếp đến là các hệ thống theo kịch bản, AIML hay các mô hình học sâu dạng seq2seq. Bên cạnh đó, các nền tảng mã nguồn mở như Rasa đã cung cấp cơ sở linh hoạt cho việc triển khai trong nhiều bối cảnh khác nhau. Mặc dù các mô hình này còn đơn giản, chúng lại giữ vai trò quan trọng trong việc xác lập nguyên tắc xử lý hội thoại, thiết kế pipeline và phương pháp đánh giá.

Ngày nay việc triển khai các hệ thống chatbot dựa trên mô hình ngôn ngữ lớn (LLM) như GPT hoặc LLaMA thường đòi hỏi phần cứng mạnh và GPU chuyên dụng, gây ra rào cản đáng kể đối với các cơ sở giáo dục địa phương có điều kiện hạ tầng hạn chế. Bên cạnh đó, các mô hình ngôn ngữ lớn thường phải trả phí khi sử dụng API.

Nghiên cứu này đề xuất một giải pháp thiết kế chatbot tư vấn học viên có yêu cầu cấu hình thấp, vận hành hiệu quả trên nền tảng CPU, đồng thời hỗ trợ tốt ngôn ngữ tiếng Việt và không tốn phí. Hệ thống được triển khai tại Trung tâm Ngoại ngữ - Tin học, Trường Đại học Kỹ thuật - Công nghệ Cần Thơ như một nghiên cứu khả thi việc kết hợp Rasa, LaBSE và xử lý ngôn ngữ tự nhiên (NLP) trong điều kiện thực tế.

Adamopoulou và Moussiades [2] cung cấp cái nhìn tổng quan về lịch sử phát triển và sự quan tâm ngày càng tăng của cộng đồng quốc

tế đối với công nghệ chatbot. Tác giả phân tích các động lực thúc đẩy việc sử dụng chatbot, làm rõ giá trị thực tiễn của chúng trong nhiều môi trường khác nhau, đồng thời đề cập đến ảnh hưởng của các định kiến xã hội trong thiết kế chatbot.

Phương pháp thực hành xây chatbot sử dụng thư viện Python và các thuật toán máy học cùng với việc tích hợp với Flask và SQLite [3] rất phù hợp với hệ thống yêu cầu phần cứng hạn chế, không sử dụng GPU.

Sharma và Joshi [4] đã thực hiện một nghiên cứu phân tích về Rasa, một nền tảng chatbot mã nguồn mở. Tính năng của lõi Rasa được nghiên cứu và ở mức độ lớn, nó có thể thực hiện các tác vụ phức tạp như tương tác với cơ sở dữ liệu, API.

Feng và cộng sự [5] đã xây dựng mô hình embedding câu đa ngôn ngữ (hơn 109 ngôn ngữ) dựa trên BERT gọi là LaBSE, giúp biểu diễn câu hiệu quả cho các tác vụ như truy xuất song ngữ (bitext retrieval), đối sánh ngữ nghĩa và truyền tải học (transfer learning). Mô hình có thể thay thế hoặc bổ sung cho LLM trong các hệ thống yêu cầu embedding ngữ nghĩa nhanh, nhẹ và đa ngữ.

Jiao [6] đã đề xuất một hệ thống chatbot thông minh sử dụng Rasa NLU kết hợp với mạng neural network (NN) để truy xuất thông tin chứng khoán (giá, vốn hóa, khối lượng giao dịch...). Kết quả nghiên cứu cho thấy Rasa NLU phù hợp cho hệ thống chatbot có số lượng dữ liệu vừa phải và yêu cầu cao về độ chính xác trong đối thoại.

Tại Việt Nam, trong lĩnh vực giáo dục số, nghiên cứu của Phương và cộng sự [7] cũng đã đề xuất xây dựng hệ thống cố vấn học tập ảo giúp trả lời tự động các câu hỏi của sinh viên về quy chế học vụ nhằm giảm tải cho giáo viên cố vấn. Bộ phân loại sử dụng các thuật toán học máy như KNN, Random Forest và SVM. Hệ thống có tiềm năng triển khai thực tế tại các trường đại học với phần cứng hạn chế (chỉ cần sử dụng CPU).

Các nghiên cứu [6][7] triển khai chatbot cho giáo dục và lĩnh vực tài chính. Tuy nhiên, các hệ thống này chưa tập trung vào tối ưu hóa CPU hoặc xử lý tiếng Việt.

2. CƠ SỞ LÝ THUYẾT VÀ PHƯƠNG PHÁP

2.1. Các mô hình chatbot

Các mô hình chatbot thường chia thành 3 nhóm chính:

- Chatbot dựa trên luật (Rule-based): Đây là thế hệ chatbot đầu tiên, hoạt động theo các kịch bản hoặc tập luật được lập trình sẵn. ELIZA là ví dụ điển hình, sử dụng quy tắc thay thế từ khóa để phản hồi người dùng. Ưu điểm là đơn giản, dễ triển khai; tuy nhiên hạn chế lớn là thiếu khả năng mở rộng và khó xử lý các tình huống hội thoại phức tạp.

- Chatbot dựa trên truy xuất (Retrieval-based): Hệ thống này lựa chọn câu trả lời từ một kho dữ liệu có sẵn dựa trên mức độ tương đồng ngữ nghĩa với câu hỏi đầu vào. Một số hệ thống sử dụng AIML hoặc kết hợp với kỹ thuật tìm kiếm và học máy để cải thiện độ chính xác. Ưu điểm là phản hồi ổn định, tránh sinh ra nội dung sai; nhược điểm là bị giới hạn trong phạm vi dữ liệu đã chuẩn bị.

- Chatbot sinh ngôn ngữ (Generative-based): Đây là loại chatbot sử dụng mô hình học máy, đặc biệt là Deep Learning (ví dụ: Seq2seq, Mạng nơ-ron phân cấp Hierarchical Neural Networks, Transformer) để tạo ra phản hồi mới thay vì chỉ chọn câu có sẵn. Loại hình này mang lại trải nghiệm tự nhiên hơn, có thể ứng phó linh hoạt với nhiều tình huống. Tuy nhiên, chúng thường yêu cầu tập dữ liệu lớn, tài nguyên tính toán mạnh và có nguy cơ sinh ra phản hồi không chính xác.

Khi học máy phát triển, các kiến trúc như Seq2seq hay Mạng nơ-ron phân cấp giúp chatbot tiến gần hơn tới hội thoại tự nhiên [8]. Những khảo sát gần đây [9][10] cũng ghi nhận xu hướng khai thác các nền tảng mã nguồn mở như Rasa để kết hợp ưu điểm của hướng rule-based và học máy. Trong công trình này,

chúng tôi kế thừa các kết quả nền tảng đó, đồng thời tích hợp LaBSE để tối ưu hiệu năng xử lý trên CPU và hỗ trợ tốt tiếng Việt.

2.2. Rasa

Rasa [11] là một hệ thống trí tuệ nhân tạo được phát triển bởi công ty Rasa sử dụng ngôn ngữ lập trình Python. Mục tiêu của Rasa là cung cấp cho các lập trình viên một công cụ mạnh mẽ với nhiều tính năng để xây dựng các ứng dụng trò chuyện tự động giữa con người và máy tính. Một trong những điểm nổi bật của Rasa là khả năng tập trung vào việc xây dựng ứng dụng chatbot dựa trên các mô hình học tăng cường (Reinforcement Learning) và học sâu (Deep Learning). Điều này cho phép Rasa tạo ra các ứng dụng có khả năng hiểu ngữ cảnh và ngữ nghĩa, tương tác với người dùng một cách tự nhiên, gần giống như cuộc trò chuyện giữa người với người.

Rasa cung cấp hai thành phần chính: Rasa NLU (Natural Language Understanding) và Rasa Core. Rasa NLU được sử dụng để hiểu và phân tích câu hỏi, phản hồi của người dùng dưới dạng cấu trúc như ý định (intent) và thực thể (entity). Trong khi đó, Rasa Core đảm nhiệm việc quản lý luồng logic trong cuộc hội thoại của ứng dụng chatbot.

Với việc cung cấp một khung làm việc mạnh mẽ và linh hoạt, Rasa đang trở thành một lựa chọn phổ biến trong việc phát triển các ứng dụng trí tuệ nhân tạo trong lĩnh vực xây dựng hệ thống trò chuyện, từ các ứng dụng thương mại cho đến các dự án cá nhân.

2.3. PostgreSQL

PostgreSQL [12] là một hệ quản trị cơ sở dữ liệu quan hệ mã nguồn mở mạnh mẽ, được sử dụng rộng rãi trong các hệ thống doanh nghiệp và học thuật. PostgreSQL hỗ trợ chuẩn SQL, tính toàn vẹn dữ liệu, đồng thời cho phép mở rộng linh hoạt thông qua các hàm, kiểu dữ liệu và mô-đun người dùng định nghĩa. Hệ thống này nổi bật với khả năng xử lý các truy vấn phức tạp, giao dịch an toàn (ACID compliant) và khả năng mở rộng theo chiều ngang và dọc. Trong nghiên cứu này, PostgreSQL được lựa chọn để lưu trữ thông tin, khóa học, lịch học... và lịch sử tương tác giữa người dùng và bot. PostgreSQL hoạt động ổn định, dễ tích hợp với Python thông qua thư viện psycopg2 và hỗ trợ tốt cho xử lý đồng thời trong môi trường chatbot hoạt động liên tục.

2.4. LaBSE

Feng và cộng sự tùy biến mô hình Multilingual BERT [13] cho 109 ngôn ngữ được gọi là LaBSE (Language-agnostic BERT Sentence Embedding). LaBSE được huấn luyện trên hàng trăm ngôn ngữ khác nhau dựa trên kiến trúc BERT và mục tiêu tối ưu ngôn ngữ đồng nhất (multilingual semantic similarity). LaBSE cho phép biểu diễn mỗi câu thành một véc-tơ 768 chiều trong không gian ngữ nghĩa chung, giúp so sánh ý nghĩa giữa các câu bằng độ đo metric cosine.

Khác với các mô hình LLM, LaBSE được thiết kế tối ưu để sử dụng suy luận trên CPU với chi phí hợp lý. Trong bài toán nhận diện ý định, việc tính véc-tơ embedding của câu hỏi và so sánh với véc-tơ câu mẫu giúp xác định ý định chính xác mà không cần huấn luyện mô hình phức tạp. Đặc biệt, LaBSE hỗ trợ rất tốt tiếng Việt và các ngôn ngữ khác.

3. KẾT QUẢ NGHIÊN CỨU

3.1. Mô hình tổng quát của chatbot

Nhóm tác giả đã xây dựng sơ đồ kiến trúc hệ thống của chatbot được trình bày như Hình 1. Trong đó:

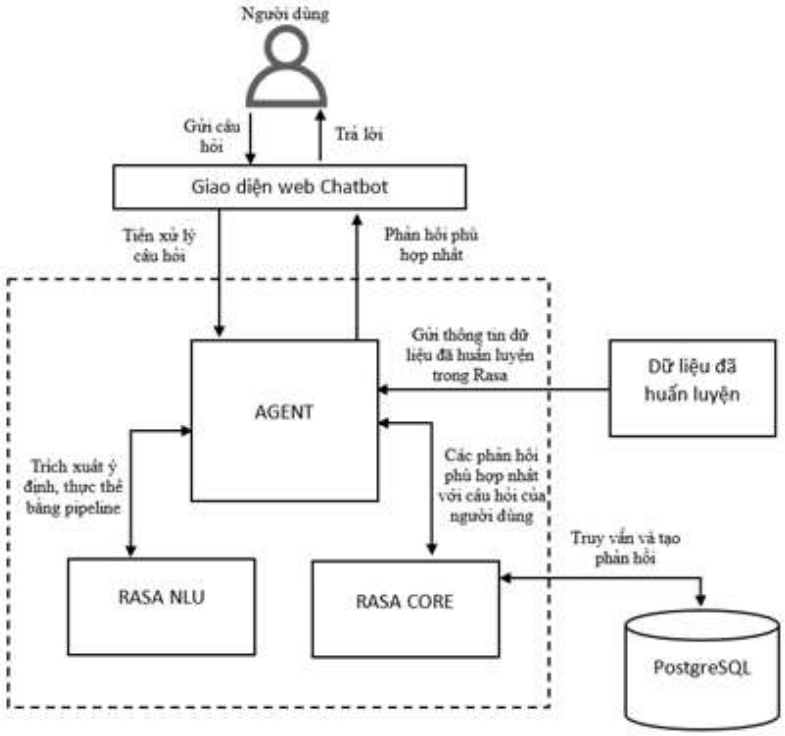
- Người dùng và giao diện web chatbot: Người dùng gửi câu hỏi qua giao diện chatbot trên nền web. Câu hỏi được tiền xử lý và chuyển đến backend xử lý ngôn ngữ.

- AGENT: Là thành phần lõi điều phối của Rasa. Nó nhận đầu vào từ giao diện, điều phối giữa các mô-đun, truy xuất dữ liệu và sinh phản hồi. AGENT hoạt động dựa trên dữ liệu huấn luyện đã được nạp vào từ trước.

- RASA NLU: Thực hiện phân tích câu hỏi bao gồm trích xuất ý định và thực thể bằng cách sử dụng pipeline NLP.

- RASA CORE: Đảm nhiệm việc chọn phản hồi thích hợp dựa trên ý định và ngữ cảnh hội thoại.

- PostgreSQL: Là nơi lưu trữ dữ liệu chatbot như lịch học, khóa học, học phí...



Hình 1. Tổng quan kiến trúc vận hành của chatbot

3.2. Quy trình tiền xử lý câu hỏi người dùng

Trước khi đưa câu hỏi người dùng vào mô hình LaBSE để sinh embedding, hệ thống thực hiện bước tiền xử lý ngôn ngữ nhằm chuẩn hóa và làm sạch câu hỏi đầu vào như sau:

- Chuẩn hóa Unicode: Chuyển văn bản về dạng chuẩn NFC để đảm bảo độ tương thích ký tự tiếng Việt.

- Loại bỏ ký tự không cần thiết: Sử dụng biểu thức chính quy để loại bỏ các ký hiệu đặc biệt, ký tự điều khiển như tab, newline.

- Chuẩn hóa câu tiếng Việt: Hiệu chỉnh các từ viết sai hoặc không chuẩn do lỗi bỏ dấu như hoà, giải,...

- Thay thế từ: Chuyển các từ viết tắt, từ đồng nghĩa về một từ chung như: chứng nhận tương đương bậc 3, b1 nội bộ, tiếng anh tương đương bậc 3…về b1 ctut bằng từ điển.

- Sửa lỗi viết sai tên khóa học: các phản hồi sai tên khóa học như: toeic ctutt, toei ctutt…được chuẩn hóa về toeic ctut bằng thư viện rapidfuzz.

- Xóa các khoảng trắng thừa.

- Tách từ: Sử dụng thư viện tách từ tiếng Việt (underthesea).

- Loại bỏ stopwords: Loại bỏ các từ không mang ý nghĩa chính như: tôi, em, con...Tập tin stopword do nhóm tác giả tự xây dựng.

- Chuyển về chữ thường: Giảm độ đa dạng dữ liệu như: Học phí, học phí, Học Phí…

Quy trình này giúp đảm bảo câu đầu vào được biểu diễn một cách nhất quán, làm tăng độ chính xác khi tính toán độ tương đồng ngữ nghĩa bằng embedding của LaBSE.

3.3. Bộ dữ liệu

Để đánh giá hiệu quả của mô hình, nhóm tác giả đã thu thập 1.712 câu hỏi từ ba nguồn (học viên đang học tại Trung tâm 30%, nhân viên tư vấn 30%, Fanpage Facebook 30% và tự thiết kế 10%). Các câu hỏi được gán nhãn intent và entity bởi cán bộ chuyên trách của Trung tâm Ngoại ngữ - Tin học, sau đó đối chiếu chéo để giảm sai lệch. Kết quả trình bày như Bảng 1.

Bảng 1. Chi tiết bộ dữ liệu hệ thống

Số câu hỏi	Ý định	Kịch bản hỏi
1.712	56	55

Về câu hỏi cho các ý định, nhóm tác giả gán thực thể để mô tả chi tiết nội dung hỏi của người dùng và xây dựng các kịch bản hỏi nhằm phản hồi thông tin hiệu quả hơn.

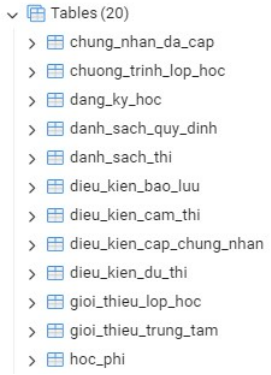
```
- intent: hoi_facebook
  examples: |
    - trung tâm có [facebook](gioi_thieu_trung_tam) không
    - tôi có thể theo dõi trung tâm trên [facebook](gioi_thieu_trung_tam) nào
    - thông tin [facebook](gioi_thieu_trung_tam) của Trung tâm
    - [facebook](gioi_thieu_trung_tam) là gì
    - tài khoản [facebook](gioi_thieu_trung_tam) của trung tâm
    - địa chỉ [facebook](gioi_thieu_trung_tam) của Trung tâm
```

Hình 2. Các câu hỏi về thông tin Facebook của Trung tâm

```
- story: Hỏi điều kiện cấp chứng nhận form
  steps:
    - intent: hoi_dieu_kien_cap_chung_nhan
    - action: form_dieu_kien_kien_cap_chung_nhan
    - active_loop: form_dieu_kien_kien_cap_chung_nhan
    - active_loop: null
    - action: action_tra_cuu_dieu_kien_kien_cap_chung_nhan
```

Hình 3. Kịch bản hỏi

Nhóm tác giả cũng xây dựng 2 cơ sở dữ liệu trong PostgreSQL bao gồm: rasa_tracker (lưu các câu hỏi và câu trả lời của người dùng và chatbot) và chatbot_cfli (lưu dữ liệu các câu trả lời của chatbot).



Hình 4. Một số bảng dữ liệu trong chatbot_c

	sender_id character varying (255)	user_message text	intent_name text	time timestamp with time zone
1	user_y60zp6uf	thông tin lịch khai giảng	{"name": "hoi_lich_khai_giang", "confidence": 1.0}	2025-06-29 09:55:14.693085+07
2	user_y60zp6uf	Bạn muốn hỏi khóa học nào? (Ví dụ: TOEIC CTUT, B1 CTUT, UD CNTT CB, ...	[null]	2025-06-29 09:55:14.741979+07
3	user_y60zp6uf	toeic ctut	{"name": "inform", "confidence": 0.8708457350730896}	2025-06-29 09:55:18.313381+07
4	user_y60zp6uf	Thông tin lịch khai giảng các lớp tiếng Anh CTUT (TOEIC CTUT) trong nă...	[null]	2025-06-29 09:55:18.511677+07
5	user_y60zp6uf	hướng dẫn đăng ký học	{"name": "hoi_dang_ky_hoc", "confidence": 1.0}	2025-06-29 09:55:27.833415+07
6	user_y60zp6uf	Bạn muốn hỏi khóa học nào? (Ví dụ: TOEIC CTUT, B1 CTUT, UD CNTT CB, ...	[null]	2025-06-29 09:55:27.885272+07
7	user_y60zp6uf	toeic ctut	{"name": "inform", "confidence": 0.8708457350730896}	2025-06-29 09:55:34.678188+07
8	user_y60zp6uf	Để đăng ký khóa học tiếng Anh CTUT (TOEIC CTUT), bạn phải <b>thực hiệ...	[null]	2025-06-29 09:55:34.774156+07

Hình 5. Câu hỏi và câu trả lời trong rasa_tracker

3.4. Pipeline

Pipeline xử lý ngôn ngữ tự nhiên tiếng Việt sử dụng DIETClassifier cho intent/entity, WhitespaceTokenizer kết hợp underthesea, CountVectorFeaturizer (n-gram), RegexFeaturizer và Sentence-Transformer LaBSE để tạo embedding 768 chiều. Pipeline trong Rasa được cấu hình với mục tiêu: hỗ trợ tiếng Việt, chạy tốt trên CPU và sử dụng mô hình LaBSE thay vì LLM. Cấu trúc pipeline được trình bày như Hình 6.

```
language: vi
pipeline:
- name: WhitespaceTokenizer
- name: LanguageModelFeaturizer
  model_name: "bert"
  model_weight: "rasa/LaBSE"
  cache_dir: null
- name: RegexFeaturizer
- name: LexicalSyntacticFeaturizer
- name: CountVectorsFeaturizer
  analyzer: char_wb
  min_ngram: 1
  max_ngram: 4
- name: DIETClassifier
  epochs: 400
  dropout: 0.3
  constrain_similarities: true
- name: EntitySynonymMapper
  epochs: 100
  constrain_similarities: true
- name: FallbackClassifier
  threshold: 0.6
  ambiguity_threshold: 0.1
```

Hình 6: Pipeline hệ thống

Pipeline đề xuất cân bằng giữa embedding mạnh (LaBSE) và đặc trưng truyền thống (n-gram, regex). Tối ưu chạy trên CPU, không yêu cầu fine-tune mô hình lớn. Phù hợp với bài toán tư vấn học viên, nơi dữ liệu không quá lớn nhưng yêu cầu chính xác cao và hỗ trợ tiếng Việt. Đối với tách từ, nhóm tác giả hiệu chỉnh WhitespaceTokenizer tách từ theo thư viện underthesea thay vì tách từ theo mặc định là khoảng trắng.

3.5. Đánh giá

Để đánh giá hệ thống, nhóm tác giả chạy câu lệnh: rasa test nlu --config configs/config.yml --cross-validation --runs 1 --folds 10 --out gridresults/config-bert và kết quả đánh giá được trình bày như Hình 7.

```
INFO     rasa.model_testing  - CV evaluation (n=10)
INFO     rasa.model_testing  - Intent evaluation results
INFO     rasa.nlu.test  - train Accuracy: 1.000 (0.001)
INFO     rasa.nlu.test  - train F1-score: 1.000 (0.001)
INFO     rasa.nlu.test  - train Precision: 1.000 (0.001)
INFO     rasa.nlu.test  - test Accuracy: 0.981 (0.010)
INFO     rasa.nlu.test  - test F1-score: 0.976 (0.014)
INFO     rasa.nlu.test  - test Precision: 0.975 (0.018)
INFO     rasa.model_testing  - Entity evaluation results
INFO     rasa.nlu.test  - Entity extractor: DIETClassifier
INFO     rasa.nlu.test  - train Accuracy: 1.000 (0.000)
INFO     rasa.nlu.test  - train F1-score: 1.000 (0.000)
INFO     rasa.nlu.test  - train Precision: 1.000 (0.000)
INFO     rasa.nlu.test  - Entity extractor: DIETClassifier
INFO     rasa.nlu.test  - test Accuracy: 0.999 (0.001)
INFO     rasa.nlu.test  - test F1-score: 0.999 (0.002)
INFO     rasa.nlu.test  - test Precision: 0.999 (0.004)
```

Hình 7. Kết quả đánh giá với cross-validation k =10

- Đối với phân loại ý định bằng 10-fold Cross Validation. Kết quả cho thấy:

+ Tập huấn luyện: F1-score: 1,000 (±0,001), Accuracy: 1,000, Precision: 1,000. Điều này cho thấy mô hình học rất tốt trên dữ liệu huấn luyện (không overfit nhờ kiểm tra).

+ Tập kiểm tra: Accuracy: 0,981 (±0,010), F1-score: 0,976 (±0,014), Precision: 0,975 (±0,018). Mức chính xác cao trên tập kiểm tra cho thấy mô hình phân loại ý định hoạt động rất tốt.

- Trích xuất thực thể bằng mô hình DIETClassifier. Kết quả cho thấy:

+ Tập huấn luyện: Accuracy/F1-score/Precision: 1,000 (±0,000). Trích xuất thực thể hoàn hảo trên dữ liệu huấn luyện.

+ Tập kiểm tra: Accuracy: 0,999, F1-score: 0,999, Precision: 0,999. Mô hình vẫn duy trì hiệu quả rất cao trên tập kiểm tra, sai số nhỏ.

Nhóm tác giả cũng đánh giá mô hình bằng ma trận nhầm lẫn.

Bên cạnh đó, nhóm tác giả thu thập 150 câu hỏi từ học viên và đánh giá out-of-scope/fallback của mô hình đề xuất. Kết quả cho thấy mô hình xác định sai ý định với tỷ lệ 10/150 (tương đương 6,7%).

Khi áp dụng cross-validation 10-fold, mô hình gốc (Baseline - CountVectorizer kết hợp Logistic) chỉ đạt Accuracy trung bình 97,3%,

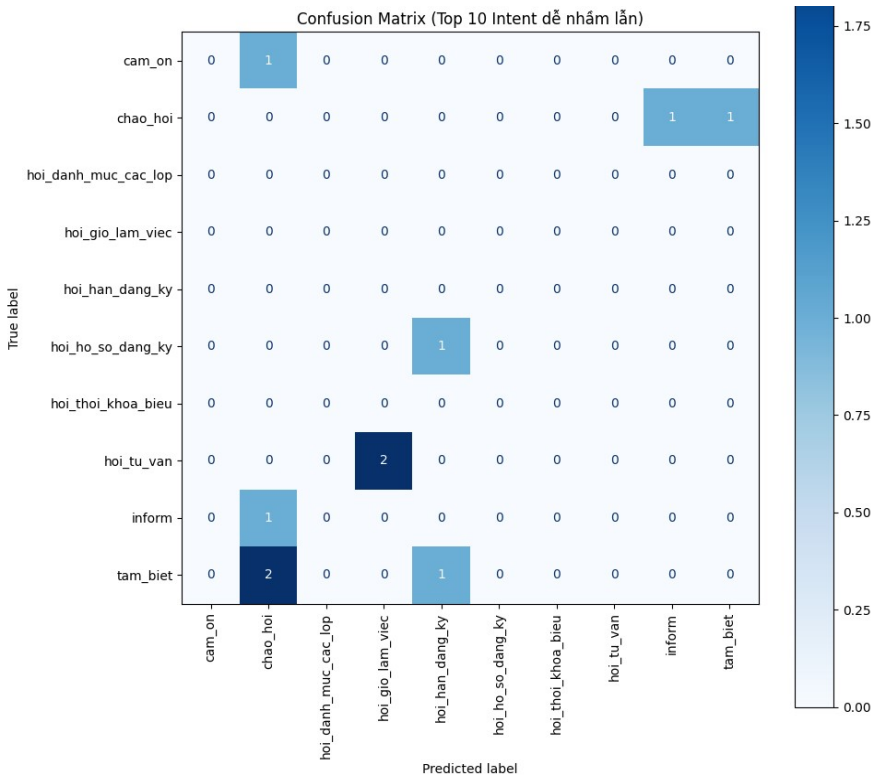
thấp hơn so với pipeline đề xuất (98,1%). Điều này chứng minh vai trò của LaBSE và bước tiền xử lý tiếng Việt.

Thử nghiệm loại bỏ từng thành phần trong pipeline (Ablation study - không dùng LaBSE, chỉ sử dụng n-gram kết hợp regex và DIET) và mô hình đề xuất cho thấy việc tích hợp embedding LaBSE giúp cải thiện chất lượng phân loại ý định. Độ chính xác tăng từ 97,5% lên 98,1% và F1-score tăng từ 0,969 lên 0,976. Với trích xuất thực thể, cả hai mô hình đều đạt kết quả tốt, nhưng mô hình có LaBSE vẫn giữ ưu thế về độ ổn định. Điều này cho thấy vai trò của LaBSE trong việc biểu diễn ngữ nghĩa tiếng Việt, giúp chatbot hiểu chính xác hơn các truy vấn có sự đa dạng về từ ngữ và ngữ cảnh.

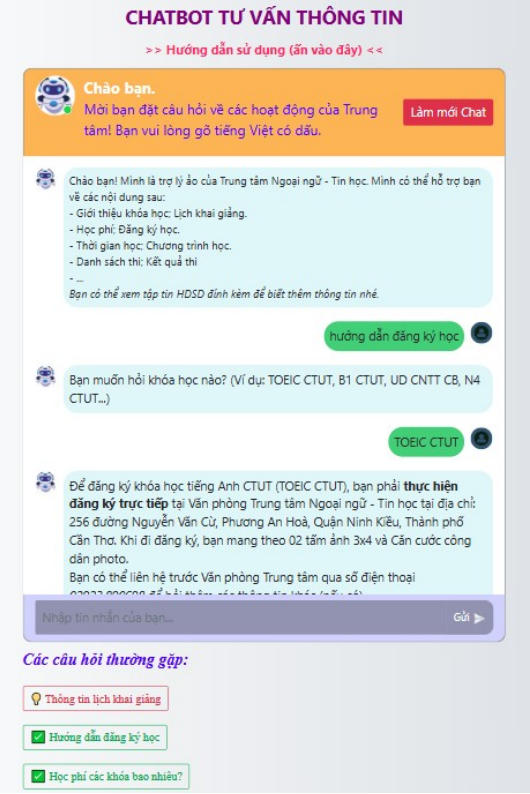
Từ các kết quả đánh giá, so sánh cho thấy mô hình đạt hiệu quả cao cả về phân loại ý định và trích xuất thực thể, khẳng định tính khả thi khi sử dụng pipeline kết hợp LaBSE trên CPU.

3.6. Triển khai thực nghiệm hệ thống

Từ các kết quả đạt được, nhóm tác giả xây dựng hệ thống bằng Flask API [14], Rasa phiên bản 3.6.21 và cài đặt trên máy chủ Nhà trường và đưa vào website của Trung tâm Ngoại ngữ - Tin học để thực nghiệm. Thời gian phản hồi trung bình 1,62 giây, độ lệch chuẩn 0,4 giây; CPU sử dụng 12%, RAM chiếm 9 GB. Đây là một kết quả rất tốt và phù hợp khi triển khai mô hình đề xuất trên máy chủ không có GPU.



Hình 8. Ma trận nhầm lẫn cho Top 10 ý định có độ nhầm lẫn nhiều nhất



Hình 9. Giao diện chatbot trên website Trung tâm Ngoại ngữ - Tin học

Để nâng cao hiệu quả phản hồi đối với câu hỏi của người dùng, nhóm tác giả cũng xây dựng tập tin hướng dẫn sử dụng chatbot bao gồm các từ khóa quan trọng, các câu hỏi thường gặp và đính kèm phía trên khung chat của chatbot.

Để nâng cấp thông tin cho chatbot, nhóm tác giả lưu nội dung truy vấn và ý định của người dùng trong cơ sở dữ liệu rasa_tracker. Các câu hỏi sẽ được đánh giá và cập nhật bổ sung vào dữ liệu hiện tại để cải tiến hiệu quả phản hồi cho chatbot.

4. KẾT LUẬN

Nghiên cứu này cho thấy tính khả thi của việc triển khai chatbot tư vấn học viên trong môi trường hạn chế tài nguyên. Điểm mới là sự kết hợp giữa Rasa và LaBSE vận hành ổn định cho CPU, không phụ thuộc GPU hoặc API thương mại. Các kết quả định lượng cho thấy độ chính xác nhận diện intent đạt 98,1%, entity đạt 99,9%, vượt trội và thời gian phản hồi dưới 2 giây đáp ứng tốt nhu cầu thực tế.

Nghiên cứu đã đề xuất pipeline dành cho CPU với LaBSE, bổ sung quy trình tiền xử lý tiếng Việt chi tiết và chứng minh hiệu quả qua các độ đo kỹ thuật. Việc tích hợp LaBSE giúp hệ thống đạt được khả năng hiểu ngữ nghĩa tốt hơn so với các phương pháp truyền thống, đồng thời hỗ trợ đa ngôn ngữ một cách hiệu quả. Bên cạnh đó, kết quả thực nghiệm cho thấy mô hình đạt độ chính xác rất cao trong cả phân loại ý định và trích xuất thực thể với tốc độ phản hồi nhanh trên phần cứng CPU thông thường. Điều này cho thấy tiềm năng ứng dụng rộng rãi trong thực tế.

Tuy nhiên, nghiên cứu vẫn còn một số hạn chế như: Dữ liệu huấn luyện còn ít (trung bình 30 câu/intent); Chưa thực hiện user study để đánh giá mức hài lòng.

Trong tương lai, nghiên cứu sẽ tập trung vào việc gợi ý câu hỏi tương tự, tích hợp với Website Trung tâm, tích hợp với Facebook Trung tâm, khảo sát sinh viên để nâng cao hiệu quả tư vấn và phản hồi thông tin về các hoạt động của Trung tâm Ngoại ngữ - Tin học.

Tài liệu tham khảo

[1] Weizenbaum, J. ELIZA—a computer program for the study of natural language communication between man and machine. Communications of the ACM. 1996; 9(1): 36–45.

[2] Adamopoulou, E. & Moussiades, L. An overview of chatbot technology. In Proceedings of the 16th International Conference on Artificial Intelligence Applications and Innovations (AIAI 2020) (pp. 373-383). Springer; 2020.

[3] Raj S. Building Chatbots with Python Using Natural Language Processing and Machine Learning. Springer Nature; 2019.

[4] Sharma R. K., Josh M. An analytical study and review of open source chatbot framework, Rasa. International Journal of Engineering Research & Technology (IJERT). 2020; 9(06).

[5] Feng F., Yang Y., Cer D., Arivazhagan N., Wang, W. (2020). Language-agnostic BERT sentence embedding. arXiv preprint , arXiv:2007.01852.

[6] Jiao A. An intelligent chatbot system based on entity extraction using RASA NLU and neural network. Journal of Physics: Conference Series. 2020; 1487(1): 012014.

[7] Nguyễn Duy Phương, Trần Việt Châu, Trần Thị Minh Thư. Cố vấn học tập ảo hỗ trợ sinh viên đại học. Kỷ yếu Hội nghị Khoa học công nghệ Quốc gia lần thứ XV về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR). Hà Nội; 2022.

[8] Serban I. V., Sordoni A., Bengio Y., Courville A. C., Pineau J. Building end-to-end dialogue systems using generative hierarchical neural network models. In Proceedings of the AAAI Conference on Artificial Intelligence. 2016; 30(1).

[9] Jia R., Xiang Z. Deep learning for chatbot: A survey. In 2017 IEEE International Conference on Computer and Information Technology (CIT) (pp. 275–282). IEEE.

[10] Pérez-Soler S., Guerra E., de Lara J. The rise of the (modelling) chatbots: Towards model-driven development of chatbots. IEEE Software. 2019; 37(5): 48–55.

[11] RASA Architecture Overview, RASA Docs. https://rasa.com/docs/rasa/arch-overview, truy cập 01/7/2025.

[12] PostgreSQL, The World's Most Advanced Open Source Relational Database. https://www.postgresql.org, truy cập 01/7/2025.

[13] Pires T., Schlinger E., Garrette, D. How multilingual is multilingual BERT?. In Proc. of the 57thAnnual Meeting of the Association for Computational Linguistics, Florence, Italy. 2019; 4996-5001.

[14] Grinberg M. Flask web development: Developing web applications with Python, 2nd ed. O'Reilly Media; 2018.

DEVELOPMENT OF AN AI CHATBOT MODEL FOR STUDENT ADVISING AT A CENTER FOR FOREIGN LANGUAGE AND INFORMATICS

ABSTRACT

Chatbots are now widely applied in education, but large language models often require high computational resources, creating challenges for local institutions. This study introduces an AI-based student counseling chatbot, experimentally deployed at the Center for Foreign Languages and Informatics, Can Tho University of Technology. The system is built on the Rasa framework combined with LaBSE and PostgreSQL, optimized to run entirely on CPUs without relying on GPUs or commercial APIs. The dataset includes 1,712 real-world questions categorized into 56 intents and 55 dialogue scenarios. Cross-validation (k=10) results show an accuracy of 98.1% for intent classification and 99.9% for entity recognition. The model demonstrates high efficiency and practical applicability in supporting student counseling within limited computing environments.

Từ khóa: *Chatbot, learner advising, rasa.*