Tạp chí Khoa học Đại học Huế: Kỹ thuật và Công nghệ; pISSN 2588-1175 | eISSN 2615-9732

Tập 131, Số 2A, 2022, Tr. 5–20; DOI: 10.26459/hueunijtt.v131i2A.6978

MỘT CẢI TIẾN CỦA PHOBERT NHẰM TĂNG KHẢ NĂNG HIỂU TIẾNG VIỆT CỦA CHATBOT THÔNG TIN KHÁCH SẠN

Ngô Văn Sơn[1]*, Nguyễn Thị Minh Nghĩa[1], Hoàng Thị Huế[1], Nguyễn Hữu Liêm[2], Võ Viết Minh Nhật[3]

[1] Trường Du lịch – Đại học Huế, Việt Nam
[2] Chi nhánh Gia Lai - Tập đoàn Bưu chính Viễn thông Việt Nam, Việt Nam
[3] Ban Đào tạo và Công tác Sinh viên – Đại học Huế, Việt Nam

Tóm tắt: Chatbot hỗ trợ thông tin du lịch bằng tiếng Việt đang thu hút nhiều sự quan tâm của giới nghiên cứu và cả kinh doanh. *Các chatbot* truyền thống *thường* được xây dựng dựa trên *cơ sở các* quy tắc, kiến thức *hay* trạng thái hữu hạn *nên thường kém hiệu quả.* Gần đây, nhờ *những bước tiến lớn của học* máy trong xử lý ngôn ngữ tự nhiên, chatbot đã đạt được những bước tiến đáng kể trong phân loại ý định, trích xuất thực thể và phân tích tình cảm. Bài báo này đề xuất một cải tiến của mô hình tiền huấn luyện nhằm xây dựng một chatbot thông tin khách sạn bằng tiếng Việt. Với đề xuất điều chỉnh các mô hình tiền huấn luyện nhằm đánh giá xem mô hình nào hoạt động tốt nhất, kết quả mô phỏng cho thấy rằng đề xuất của chúng tôi đã mang lại hiệu quả đáng kể với Accuracy 96,4%, F1-score 96,9% và Precision 97,4%.

Từ khóa: chatbot, xử lý ngôn ngữ tự nhiên, học máy, mô hình tiền huấn luyện, độ đo

An improvement of PhoBERT to increase the Vietnamese understanding of the hotel information chatbot

Ngo Van Son[1]*, Nguyen Thi Minh Nghia[1], Hoang Thi Hue[1], Nguyen Huu Liem[2], Vo Viet Minh Nhat[3]

[1] School of Hospitality & Tourism – Hue University, Vietnam
[2] Vietnam Posts and Telecommunications Group – Gia Lai Branch, Vietnam
[3] Department of Academic and Students' Affairs – Hue University, Vietnam

Abstract. Chatbots supporting tourism information in Vietnamese are attracting a lot of attention from researchers and businesses alike. Traditional chatbots are often built on the basis of finite rules, knowledge, or states, so they are often ineffective. Recently, thanks to the implementations of machine learning in natural language processing, chatbots have made significant strides in intent classification, entity extraction, and sentiment analysis. This paper proposes an improvement of the pre-training model to build a hotel information chatbot in

** Liên hệ: ngovanson@hueuni.edu.vn*

Nhận bài: 27–10–2022; Ngày nhận đăng: 08–11–2022

Vietnamese. With the suggestion of adjusting the pre-trained models to evaluate which model works best, the simulation results show that our proposal obtained a significant effect, based on the metric for evaluating, namely, Accuracy = 96.4%; F1-score = 96.9%; and Precision = 97.4%.

Keywords: chatbot, natural language processing, machine learning, pre-trained models, metric

1 Đặt vấn đề

Chatbot hay trợ lý ảo xuất hiện ngày càng nhiều trong đời sống của con người như Apple Siri[1], Google Assistant[2], Amazon Alexa[3], Microsoft Cortana[4] hay FPT AI Chat[5]. Với sự bùng nổ của ứng dụng công nghệ thông tin vào hoạt động kinh doanh du lịch, việc khách hàng tiếp cận thông tin du lịch qua chatbot trở nên phổ biến. Người dùng có thể giao tiếp với chatbot thông qua tương tác bằng giọng nói hoặc văn bản. Chatbot có thể giúp người dùng thực hiện nhiều yêu cầu khác nhau từ gọi điện, tìm kiếm thông tin đến đặt vé, đặt phòng khách sạn. Các chatbot truyền thống được xây dựng dựa trên các quy luật, thói quen trong ngôn ngữ của người dùng nên khi gặp các câu hỏi ngoài kịch bản, chatbot sẽ không hiểu và phải đợi nhân viên chăm sóc khách hàng can thiệp. Có hai mô hình chatbot truyền thống: Chatbot theo kịch bản (Menu/Button), như Chatbot Messnow[6] và Bot Bán hàng[7] và Chatbot nhận dạng từ khoá như Chatbot Harafunnel[8]. Ưu điểm của hai loại chatbot truyền thống này là dễ xây dựng và độ chính xác cao vì người dùng đưa ra yêu cầu dựa trên những gợi ý (nút) đã được xây dựng trước. Tuy nhiên, người dùng sẽ bị động trước những mong muốn của mình vì phụ thuộc vào các lựa chọn được cung cấp của chatbot. Mô hình thứ hai có nhiều ưu điểm hơn so với mô hình thứ nhất vì nó cho phép người dùng chủ động hơn trong việc đưa ra yêu cầu, nhưng nếu người dùng sử dụng các từ đồng nghĩa với các từ khoá thì chatbot không thể hiểu để trả lời một cách phù hợp, cũng như không nắm bắt được ngữ cảnh cuộc trò chuyện.

Để cải thiện khả năng phục vụ, chatbot ngày nay sử dụng các kỹ thuật tiên tiến về xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP), các kỹ thuật học sâu như Deep

[1] https://www.apple.com/siri

[2] https://assistant.google.com

[3] https://developer.amazon.com/alexa

[4] https://www.microsoft.com/en-us/cortana

[5] https://fpt.ai/vi/chat-bot-vi

[6] https://messnow.com

[7] https://botbanhang.vn

[8] https://en.harafunnel.com

Reinforcement Learning (DRL) hay Deep Neural Network (DNN) [12]. Gần đây, cộng đồng xử lý ngôn ngữ tự nhiên đã đạt được nhiều đột phá trong việc tích hợp ngữ cảnh và các mô hình ngôn ngữ hai chiều như ELMo [11], OpenAI GPT-2 [14], BERT 5, RoBERTa [6], DistilBERT [15], XLM [1] và XLNet [22]. Đặc biệt, mô hình BERT là một tiếp cận xử lý ngôn ngữ tự nhiên tiêu biểu với các khả năng dịch ngôn ngữ, phân loại câu, v.v.

Đã có nhiều nghiên cứu xây dựng chatbot cho ngành du lịch, trong đó đa số xem xét từng trường hợp dữ liệu cụ thể bằng ngôn ngữ tiếng Anh áp dụng cho ngành hàng không [8], đại lý du lịch [13], lập kế hoạch chuyến đi [16], hỗ trợ thông tin du lịch tại điểm đến [7,20], v.v. Cũng không có khẳng định nào về mô hình học máy tốt nhất cho mọi trường hợp. Trên miền dữ liệu tiếng Việt, vấn đề bóc tách từ ghép hiện nay đã có những nghiên cứu giải quyết như thư viện UnderTheSea của Vu Anh [21]. Bên cạnh đó, các vấn đề về chuẩn hoá các lỗi sai chính tả, viết tắt, v.v. đã ảnh hưởng rất lớn đến độ chính xác của mô hình.

Trong bài báo này, chúng tôi đề xuất một cải tiến đối với PhoBERT [3] – một mô hình ngôn ngữ được tiền huấn luyện trên dữ liệu dành riêng cho tiếng Việt. Cải tiến này giúp chatbot có thể phân loại ý định và trích chọn thông tin tốt hơn khi phân tích câu trong miền tiếng Việt. Mục tiêu của cải tiến là hiểu chính xác ý định của du khách.

Các đóng góp chính của bài báo bao gồm:

– Mô hình hoá bài toán hiểu tiếng Việt cho chatbot thông tin khách sạn dựa trên học máy (Machine Learning-Chatbot – ML-Chatbot), từ đó có thể áp dụng vào nhiều miền ứng dụng khác nhau như quán café, nhà hàng, đặt vé trực tuyến và đại lý du lịch.

– Cải tiến mô hình tiền huấn luyện PhoBERT nhằm tăng năng lực hiểu tiếng Việt của ML-Chatbot.

Nội dung tiếp theo của bài báo bao gồm: Mục 2 tóm lược và đánh giá các nghiên cứu liên quan trong bốn năm trở lại đây, trong đó tập trung vào các mô hình ứng dụng học máy vào chatbot sử dụng tiếng Việt, đặc biệt áp dụng trong lĩnh vực du lịch. Trên cơ sở các phân tích, Mục 3 sẽ mô tả chi tiết các bước xây dựng mô hình chatbot thông tin du lịch với khả năng hiểu tiếng Việt. Các bước xây dựng bao gồm: chuẩn bị và chuẩn hóa dữ liệu cho việc huấn luyện, làm rõ mô hình chatbot, xác định phương pháp đánh giá và xây dựng kịch bản thử nghiệm. Các phân tích về đánh giá độ chính xác hiểu ngôn ngữ tự nhiên của chatbot được mô tả ở Mục 4. Cuối cùng, kết luận được trình bày trong Mục 5.

2 Các nghiên cứu liên quan

Trong thập kỷ đầu tiên của thế kỷ 21, những đột phá trong lĩnh vực học máy đã thúc đẩy chatbot phát triển. Bài báo này tập trung vào việc đánh giá các mô hình chatbot với khả năng hiểu ngôn ngữ tự nhiên từ năm 2018 đến nay.

Trieu Hai Nguyen và cs. [19] đã đào tạo lại PhoBERT để trích xuất các đặc điểm của dữ liệu văn bản. Nghiên cứu sử dụng thuật toán phân cụm K-Means và DBSCAN cho các nhiệm vụ phân nhóm dựa trên các nhúng đầu ra từ PhoBERT. Kết quả cho thấy mô hình PhoBert hoạt động hiệu quả hơn. Tuy nhiên, kết quả nghiên cứu giải quyết bài toán phân cụm chưa làm rõ cụ thể các bài toán phân loại ý định, trích xuất thực thể dành cho chatbot.

Trong nghiên cứu của Bozic, Tazl và Wotawa [4], nền tảng Dialogflow được sử dụng để xây dựng kế hoạch tương tác của chatbot, trong đó mỗi hành động có thể được giả định là một câu hỏi được đưa ra cho chatbot. Câu trả lời của chatbot phải làm cho điều kiện hậu hành động trở thành thực tế để thực hiện kế hoạch. Trong trường hợp có sự sai lệch giữa hành vi chatbot thực tế và hành vi dự kiến, cần phải lập kế hoạch lại. Với cách tiếp cận này, nghiên cứu cũng đã thử nghiệm ứng dụng vào lĩnh vực du lịch. Tuy nhiên, chatbot này còn nhiều hạn chế cần chỉnh sửa ở phần kế hoạch tương tác.

Nguyễn Thanh Thủy [8] đã phân tích việc xác định ý định của người dùng đóng vai trò quan trọng như thế nào trong thiết kế hệ thống chatbot bởi vì nó sẽ quyết định đến câu trả lời hay hành vi kế tiếp của chatbot. Nghiên cứu này đã đề xuất giải pháp ứng dụng thuật toán học có giám sát Multi-Class SVM (Support Vector Machine) để xây dựng hệ thống chatbot hỏi – đáp tiếng Việt, sao cho nó có thể giúp chatbot hiểu và giao tiếp với con người thông qua đàm thoại văn bản. Nghiên cứu đã sử dụng kỹ thuật túi từ (BoW – Bag of Words) kết hợp với phương pháp TF-IDF (Term Frequency – Inverse Document Frequency) để xây dựng véc-tơ đặc trưng ngữ nghĩa của các câu văn bản tiếng Việt, sử dụng thuật toán Multi-Class SVM để huấn luyện và phân lớp, sau đó so sánh độ chính xác với các thuật toán khác. Hệ thống chatbot này đã được mô phỏng để trả lời tự động một số câu hỏi thường gặp của khách hàng khi sử dụng dịch vụ của Vietnam Airlines. Các chỉ số Accuracy, Macro-average Precision, Macro-average Recall và Macro-average F1-Score đã được sử dụng để so sánh và đánh giá các mô hình. Kết quả mô phỏng cho thấy giải thuật SVM có độ Accuracy tốt nhất (0.87429) và F1-Score nhỉnh hơn so với giải thuật NBs (0.05–0.001) và trội hơn nhiều so với giải thuật kNN (0.11–0.103) và DT (0.24–0.222).

Le và cs. [9] đã xây dựng bộ lọc đánh giá phản hồi của chatbot dựa trên đặc điểm của ngôn ngữ tiếng Việt. Nếu phản hồi không phù hợp, hệ thống sẽ không trả lời người dùng. Trong trường hợp này, hệ thống có thể yêu cầu thêm thông tin để tìm cách phản hồi tốt hơn hoặc ít nhất là trả lại thông báo thay vì phản hồi không chính xác. Nghiên cứu này đã phát triển một chatbot trên nền web nhằm đánh giá cách tiếp cận của họ.

Trong nghiên cứu của mình, Nguyen và Shcherbakov [18] đã triển khai một chatbot tiếng Việt có khả năng hiểu ngôn ngữ tự nhiên, có thể tạo phản hồi, thực hiện hành động với người dùng và ghi nhớ bối cảnh của cuộc trò chuyện. Các tác giả đã sử dụng nền tảng RASA để xây dựng chatbot và đề xuất một cách tiếp cận bằng cách sử dụng cấu hình tùy chỉnh cho mô hình hiểu ngôn ngữ tự nhiên (Natural Language Understanding – NLU). Nghiên cứu này áp dụng ba mô hình gồm: hai mô hình tiền huấn luyện là FastText [2] và BERT đa ngôn ngữ; một mô hình tùy chỉnh không sử dụng mô hình tiền huấn luyện. Để đánh giá và so sánh mô hình đề xuất, các tác giả đã sử dụng Accuracy, F1-score và Precision. Kết quả thực nghiệm so sánh ba mô hình cho thấy rằng mô hình được đề xuất hoạt động tốt hơn trong phân loại ý định và trích xuất thực thể.

Oanh Thi Tran và Tho Chi Luong [10] đã đề xuất một khung mô hình hóa bài toán phân loại và bài toán ghi nhãn trình tự hai lớp dựa trên mô hình học sâu. Hệ thống chatbot có thể phát hiện ý định và nhận ra ngữ cảnh hội thoại trong miền thương mại điện tử Việt Nam nhằm giúp các thương hiệu bán lẻ giao tiếp tốt hơn với khách hàng của họ. Kết quả thử nghiệm của nghiên cứu cho thấy rằng mạng no-ron học sâu có thể hoạt động tốt hơn các phương pháp học máy thông thường. Trong việc phát hiện ý định, nghiên cứu cho thấy số đo F tốt nhất là 82,32%. Trong việc trích xuất các ngữ cảnh, phương pháp đề xuất có số đo F nằm trong khoảng từ 78 đến 91% tùy thuộc vào các loại ngữ cảnh cụ thể.

Tóm lại, các nghiên cứu nêu trên đều sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên, các kỹ thuật học sâu, kỹ thuật túi từ, kết hợp tần suất xuất hiện từ và các mô hình ngôn ngữ hai chiều. Tùy thuộc vào từng trường hợp cụ thể mà mỗi nghiên cứu chỉ ra mô hình nào là phù hợp nhất. Trong bài báo này, chúng tôi tiếp tục khảo sát hiệu quả phân loại ý định và trích xuất thực thể đối với mô hình BERT cho tiếng Việt, cụ thể là PhoBERT và so sánh với mô hình FastText và BERT có hỗ trợ tiếng Việt. Dựa vào kết quả đánh giá, chúng tôi ứng dụng vào việc xây dựng chatbot thông tin khách sạn.

3 Mô hình ML-Chatbot thông tin khách sạn

3.1 Mô hình chatbot dựa trên học máy

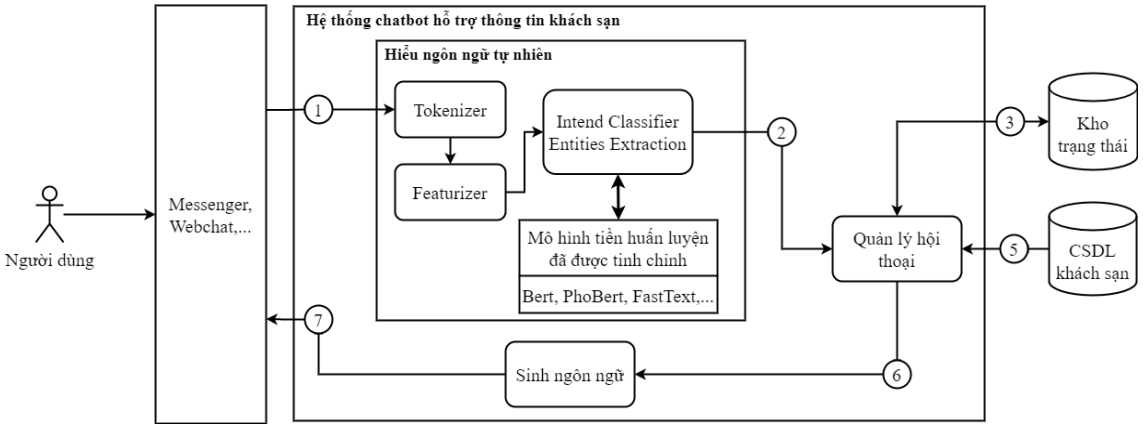
Các mô hình chatbot hiện nay tương tác với người dùng thông qua âm thanh hoặc văn bản. Có nhiều nền tảng khác nhau như Google Dialogflow[9] hay Microsoft bot framework[10] để lựa chọn xây dựng chatbot, nhưng chúng tôi chọn RASA[11] vì nó là một khung làm việc mã nguồn mở và có thể tùy chỉnh khá linh hoạt so với các nền tảng khác. RASA được thiết kế để người dùng có

[9] https://dialogflow.cloud.google.com

[10] https://dev.botframework.com

[11] https://rasa.com

thể phát triển dễ dàng với nhiều khả năng tùy chỉnh được cung cấp như thêm thành phần tùy chỉnh linh động theo trường hợp cụ thể hoặc thêm nhiệm vụ hoặc thêm nhiều webhook cho một dự án. Trên cơ sở nền tảng RASA, chúng tôi đã mô hình hóa chatbot dựa trên học máy (ML-Chatbot) áp dụng cho trường hợp hỗ trợ thông tin khách sạn.



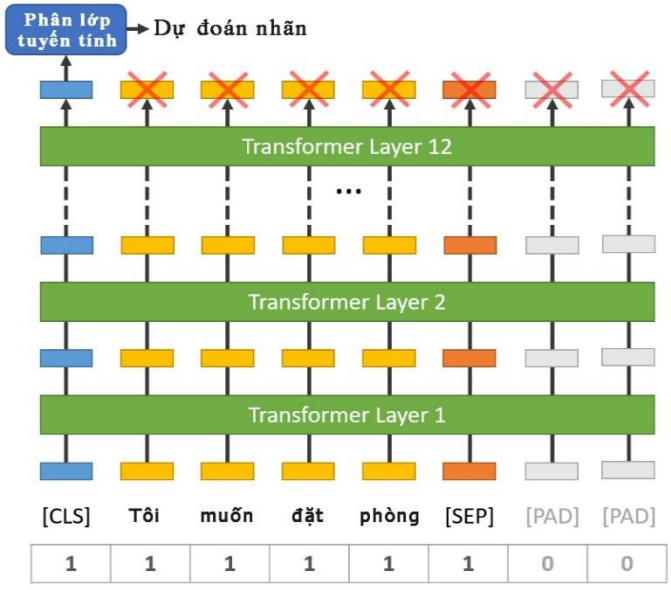
Hình 1. Mô hình ML-Chatbot tương tác bằng văn bản

Mô hình ML-Chatbot, hiểu ngôn ngữ tự nhiên, có nhiệm vụ phân tích cú pháp của người dùng và đưa chúng thành một biểu diễn ngữ nghĩa có cấu trúc. Biểu diễn này thường có dạng là *cặp ý định* (intent) và cặp *thuộc tính – giá trị* (slot: value). Ý định chỉ chức năng của lời nói như truy vấn hoặc cung cấp thông tin. Cặp *thuộc tính – giá trị* là yếu tố ngữ nghĩa được đề cập trong câu nói. Ví dụ, trong câu nói "Bạn có thể giới thiệu một nhà hàng Hàn Quốc ở Huế không?", các cặp *thuộc tính – giá trị* có thể là ("ẩm thực" – "Hàn Quốc") và ("địa điểm" – "Huế"). Ý định ở đây là "cung cấp thông tin". Phát hiện ý định (intent detection) và trích xuất *thuộc tính – giá trị* (slot-value extraction) có thể được giải quyết bằng cách sử dụng mạng nơ-ron hồi quy (RNN), mạng nơ-ron tích chập (CNN), mạng nơ-ron đệ quy, CRF hoặc mô hình BERT.

Trong mô hình còn có các thành phần truy xuất và lưu trữ thông tin nhằm hỗ trợ du khách. Ví dụ, nếu du khách cần đặt phòng tại khách sạn thì hệ thống xác nhận chatbot đang trong ngữ cảnh muốn đặt phòng khách sạn nhưng chưa rõ cần đặt phòng loại nào nên chatbot sẽ đưa ra quyết định hỏi lại người dùng. Trong trường hợp mà chatbot có đầy đủ thông tin hỏi về đặt phòng thì sẽ lấy dữ liệu từ Cơ sở dữ liệu khách sạn để trả về cho người dùng. Kho trạng thái giúp lưu ngữ cảnh và trạng thái hội thoại hiện tại. Việc lưu trạng thái các hội thoại giúp chatbot kiểm soát được nhiều hội thoại của nhiều khách hàng đồng thời. Mặc khác, cơ sở dữ liệu khách sạn chứa các thông tin cụ thể về khách sạn và các dịch vụ liên quan. Thông tin này được cập nhật thường xuyên, đúng với thực tế của khách sạn. Với cơ sở dữ liệu này, việc thay đổi các thông tin về khách sạn sẽ không cần cập nhật và huấn luyện lại ML-Chatbot.

3.2 Cải tiến PhoBERT

Vấn đề quan trọng nhất của chatbot được thiết kế để thực hiện các tác vụ cụ thể trong một lĩnh vực nhất định như đặt phòng khách sạn, mua vé máy bay và cung cấp thông tin về đại dịch Covid-19 là dữ liệu khó đảm bảo đủ lớn để huấn luyện từ ban đầu. Quá trình huấn luyện một mô hình học máy trên bộ dữ liệu này dẫn tới kết quả không thực sự tốt và lãng phí tài nguyên tính toán. Do vậy, trong phần hiểu ngôn ngữ tự nhiên của chatbot cần hướng đến tinh chỉnh các mô hình tiền huấn luyện nhằm giúp chatbot hiểu được ngôn ngữ tự nhiên theo đúng miền ứng dụng cụ thể. Phương pháp này được xây dựng dựa trên ý tưởng chuyển giao tri thức đã được học từ những mô hình tốt trước đó để có được mô hình tận dụng lại các tri thức đã tiền huấn luyện đó. Việc tận dụng lại tri thức từ các mô hình tiền huấn luyện với cùng tác vụ sẽ giúp các mô hình được huấn luyện dự báo tốt hơn với dữ liệu mới vì mô hình được học trên cả hai nguồn tri thức đó là dữ liệu huấn luyện và dữ liệu mà nó đã được học trước đó.



Hình 2. Mô hình tinh chỉnh PhoBERT

Để cải tiến PhoBERT, chúng tôi chọn tinh chỉnh nhiệm vụ phân loại ý định người dùng. PhoBERT là một mô hình tiền huấn luyện, mã hóa văn bản từ ngôn ngữ tiếng Việt thành một không gian nhúng chung. Đây là mô hình tiền huấn luyện dựa trên kiến trúc cải tiến RoBERT so với BERT và được huấn luyện trên khoảng 20 GB dữ liệu bao gồm khoảng 1 GB Vietnamese Wikipedia corpus và 19 GB còn lại lấy từ Vietnamese news corpus. Mô hình PhoBERT cũng có hai phiên bản PhoBERT-Base với 12 lớp và PhoBERT-Large với 24 lớp. Để cải tiến mô hình trong bài toán phân lớp, các tham số trong mô hình PhoBERT cần được chỉnh sửa nhằm thực hiện phân loại ý định. Sau đó, tiếp tục huấn luyện mô hình trên tập dữ liệu khách sạn cho đến khi toàn bộ

mô hình phù hợp với nhiệm vụ phân loại ý định. Để mô hình sau khi tinh chỉnh đáp ứng được nhiệm vụ phân loại ý định, chúng tôi thêm một lớp phân loại ở cuối và đầu ra của PhoBERT tinh chỉnh sẽ là đầu vào của lớp phân loại này. Đầu vào của PhoBERT tinh chỉnh cũng phải được thay đổi để phù hợp với nhiệm vụ phân lớp. Theo mô hình PhoBERT, văn bản đầu vào của mô hình phải được chuyển thành chuỗi token[12] và được chèn thêm hai token [CLS] và [SEP]. Trong nhiệm vụ phân lớp, trạng thái ẩn tương ứng với token đặc biệt [CLS] là đại diện của toàn bộ câu được sử dụng cho các nhiệm vụ phân loại, khác với vectơ trạng thái ẩn tương ứng với token biểu diễn từ thông thường.

Như vậy, khi cung cấp một câu đầu vào cho mô hình trong quá trình huấn luyện, đầu ra là một véc-tơ trạng thái ẩn tương ứng với token này. Lớp bổ sung được thêm ở trên bao gồm các nơ ron tuyến tính chưa được huấn luyện có kích thước [*kích thước véc-tơ trạng thái ẩn, số ý định*], có nghĩa là đầu ra của PhoBERT kết hợp với lớp phân loại là một vectơ gồm hai số đại diện cho điểm số để làm cơ sở phân loại câu. Vectơ này được đưa tiếp vào hàm mất mát cross-entropy để tính phân bố xác suất phân loại ý định. Trong bài báo này, chúng tôi chọn các tham số được sử dụng trong việc tinh chỉnh mô hình PhoBERT như sau:

- max_length=128; //chiều dài tối đa của văn bản đưa vào, tính theo số token.

- Mô hình BERT mặc định áp dụng cho dịch ngôn ngữ. Đầu vào được chèn thêm [SEP] để phân tách hai câu. Tuy nhiên, bài toán phân loại chỉ cần đầu vào một câu và để đảm bảo cấu trúc đầu vào của mô hình, chúng tôi thiết lập tham số sau đây nhằm tự động thêm khoảng đệm vào phía sau [SEP]: pad_to_max_length=true;

- learning_rate (adam)=1e-5; //tỷ lệ học theo phương pháp tối ưu Adam

- epochs=20; //số lần đưa toàn bộ dữ liệu vào huấn luyện

- batch_size=8; //thiết lập xác định toàn bộ dữ liệu được tách thành bao nhiêu

- cross-entropy; //hàm mất mát được sử dụng là cross-entropy

Mô hình PhoBERT sau khi tinh chỉnh được lưu thành định dạng chuẩn của Huggingface[13] để đưa vào sử dụng trong ML-Chatbot.

3.3 Cấu hình ML-Chatbot

Cấu hình là bước cần thiết để các thành phần phối tự động tạo ra một mô hình học máy. Đối với các hệ thống chatbot sử dụng mô hình học máy thì cấu hình này là phần đóng gói toàn

[12] Token là các khối xây dựng trong xử lý ngôn ngữ tự nhiên. Token có thể là một từ (word), một từ phụ (sub-word) hoặc một ký tự (character).

[13] https://huggingface.co

bộ các phương pháp xử lý dữ liệu để tạo ra một mô hình học máy phù hợp nhất cho một bộ dữ liệu cụ thể. Trong bước cấu hình hoạt động ML-Chatbot, chúng tôi sử dụng thêm hai mô hình tiền huấn luyện sau đây nhằm đánh giá và so sánh với mô hình tinh chỉnh PhoBERT:

(1) Với mô hình BERT, chúng tôi chọn bert-base-multilingual-cased[14] được lưu trữ trên thư viện tiền huấn luyện Huggingface có thể sử dụng với 12 lớp ẩn được luấn luyện trên 104 ngôn ngữ (bao gồm tiếng Việt) và phân biệt chữ hoa và chữ thường.

(2) Mô hình FastText do Facebook phát triển theo phương pháp n-gram (tần suất xuất hiện của *n* kí tự liên tiếp xuất hiện trong dữ liệu), huấn luyện trên 157 ngôn ngữ. Chúng tôi tải thư viện cc.vi.300.bin[15] (hơn 7 GB) tiền huấn luyện mô hình FastText dành cho ngôn ngữ tiếng Việt để sử dụng trong cấu hình hoạt động của ML-Chatbot.

Bảng 1. Cấu hình điều chỉnh mô hình

Cấu hình	Mô hình	Tokenizer	Featurizer	Phân loại	Trích xuất thực thể
BERT	bert-base-multilingual-cased	WordPiece [5]	Theo mô hình ngôn ngữ	Sklearn	CRF
FastText	FastText (cc.vi.300.bin)	UETsegmenter	n-gram	Sklearn	CRF
PhoBertBase	phoBERT-base tinh chỉnh	VnCoreNLP [3, 17]	Theo mô hình ngôn ngữ	Sklearn	CRF

Đối với mô hình FastText, bộ tách từ UETsegmenter được sử dụng mặc định để tách từ trong ngôn ngữ tiếng Việt. Trong khi đó bert-base-multilingual-cased sử dụng phương pháp tách từ WordPiece và trong PhoBERT việc tách từ được sử dụng theo thư viện VnCoreNLP.

3.4 Dữ liệu

Dữ liệu sử dụng trong việc nâng cao năng lực hiểu ngôn ngữ tiếng Việt của ML-Chatbot được thu thập từ dữ liệu hỏi đáp về thông tin du lịch. Sau đó, chúng tôi biên tập thành bộ dữ liệu câu hỏi – trả lời theo từng chủ đề liên quan đến nhu cầu cần thiết của khách đi du lịch như ăn, ở, đi lại và thông tin về các địa danh. Trong quá trình thu thập dữ liệu thông tin, chúng tôi đã khai thác từ các phương tiện thông tin đại chúng như các báo điện tử, các trang website du lịch,

[14] https://huggingface.co/bert-base-multilingual-cased

[15] https://fasttext.cc/docs/en/crawl-vectors.html

TripAdvisor.com, booking.com và locatravel. Ngoài ra, dữ liệu cũng được thu thập từ các nhân viên chăm sóc khách hàng trực tuyến tại khách sạn Thanh Lịch, Huế. Bên cạnh đó, chúng tôi còn thu thập dữ liệu từ những phần hỏi đáp, đánh giá và phản hồi của một số website khách sạn khác. Chúng tôi cũng tiến hành đi thực tế ở nhiều địa danh du lịch nổi tiếng ở Huế như Đại nội, Chùa Thiên Mụ và các lăng tẩm để tìm hiểu sâu hơn về nhu cầu của khách du lịch như về giá dịch vụ tham quan, các dịch vụ về phương tiện di chuyển và hiểu thêm về giá cả và chất lượng của các dịch vụ. Dựa trên dữ liệu đã thu thập, chúng tôi tiến hành phân tích, trích lọc và xây dựng tập huấn luyện và kịch bản hội thoại. Đây là phần khá tốn thời gian vì cần xác định các nhóm câu hỏi giống nhau để phân loại và câu trả lời cần được thống kê lại theo nhóm nhằm xây dựng cơ sở dữ liệu hỗ trợ chatbot truy vấn tìm thông tin.

Việc xây dựng cơ sở dữ liệu cũng giúp cho quá trình cập nhật dữ liệu mới và chatbot sẽ phân biệt được dữ liệu mới nhất với dữ liệu trước đây. Sau khi qua bước tiền xử lý thủ công, dữ liệu huấn luyện có dạng như sau:

Dữ liệu về ý định và các ví dụ liên quan đến ý định:

- intent: book_room_with_num_people

examples: |

- Tôi muốn đặt phòng cho [2]{"entity": "num_people", "value": "2"} người

- Tôi muốn đặt phòng cho [3]{"entity": "num_people", "value": "3"} người

- Tôi muốn đặt phòng cho [5]{"entity": "num_people", "value": "5"} người

- Tôi muốn đặt phòng cho [6]{"entity": "num_people", "value": "6"} người

- Tôi muốn đặt phòng cho [10]{"entity": "num_people", "value": "10"} người

- Tôi muốn đặt phòng cho [20]{"entity": "num_people", "value": "20"} người

- intent: book_room_with_number

examples: |

- Tôi muốn đặt [hai]{"entity": "num_rooms", "value": "2"} phòng

- Tôi muốn đặt [ba]{"entity": "num_rooms", "value": "3"} phòng

- Tôi muốn đặt [hai]{"entity": "num_rooms", "value": "2"} phòng cho gia đình mình

- Tôi muốn đặt [2]{"entity": "num_rooms", "value": "2"} phòng để ở

- Tôi muốn [3]{"entity": "num_rooms", "value": "3"} phòng

- Tôi muốn đặt [34]{"entity": "num_rooms", "value": "34"} phòng.

- Tôi muốn đặt [8]{"entity": "num_rooms", "value": "8"} phòng.

Dữ liệu về câu chuyện hội thoại, kịch bản tương tác giữa khách và chatbot:

```
stories:
  - story: book room
    steps:
      - intent: greet
      - action: utter_greet
      - intent: book_room
      - action: booking_form
      - active_loop: booking_form
      - slot_was_set:
          - num_rooms: 2
          - room_type: "Simple"
```

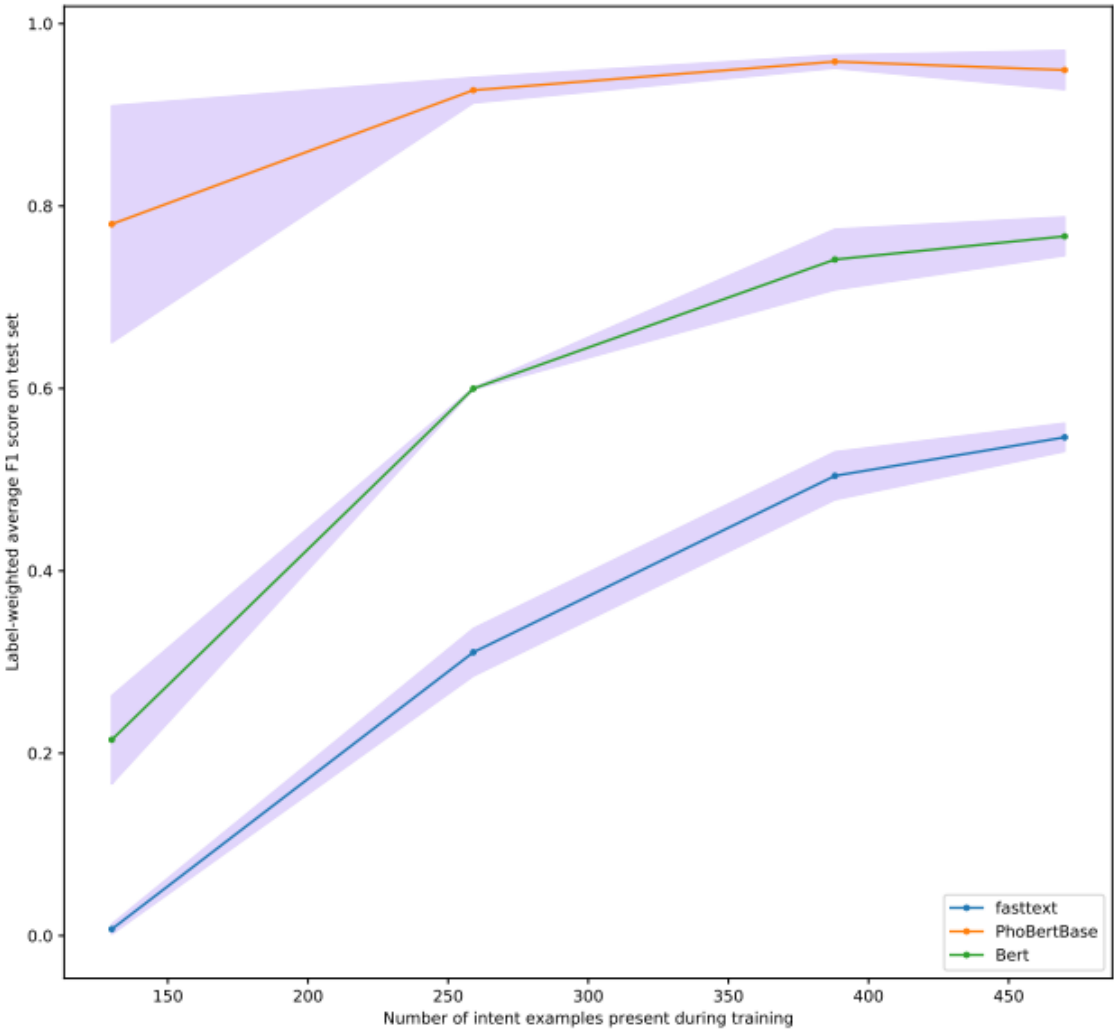
Kết quả dữ liệu sau thu thập bao gồm hơn 700 mẫu câu phân loại ý định, 59 ý định và 8 thực thể liên quan đến thông tin cơ bản về khách sạn.

4 Thử nghiệm và phân tích kết quả

Chúng tôi sử dụng RASA phiên bản 2.8.2 để tiến hành thử nghiệm mô hình với dữ liệu huấn luyện được thu thập từ các trang hỏi đáp của khách sạn. Dữ liệu này được biên tập lại theo cấu trúc của RASA với 59 ý định, 8 thực thể và hơn 700 mẫu câu. Các cài đặt được thực hiện trên máy tính PC (hệ điều hành Windows 10) với cấu hình 2.7 GHz Intel Core 4 CPU, 8G RAM.

Với tập dữ liệu nhỏ hơn 700 mẫu, để tránh hiện tượng overfitting và underfitting khi xây dựng mô hình, chúng tôi đã sử dụng kỹ thuật Leave-One-Out (một trường hợp của k-Fold cross validation) để tổ chức tập huấn luyện và tập kiểm tra trong quá trình huấn luyện, đánh giá mô hình.

Hình 3 cho thấy sự thay đổi của mô hình được đánh giá theo chỉ số F1 với số lượng mẫu tăng dần. Phân tích kết quả, chúng tôi có thể kết luận rằng PhoBERT hoạt động vượt trội hơn so với BERT và FastText đối với bộ tiền huấn luyện mô hình ngôn ngữ tiếng Việt. Ngoài ra, kết quả cũng cho thấy mô hình PhoBERT và BERT (với cấu hình bert-base-multilingual-cased) hiệu quả hơn nhiều so với mô hình FastText.



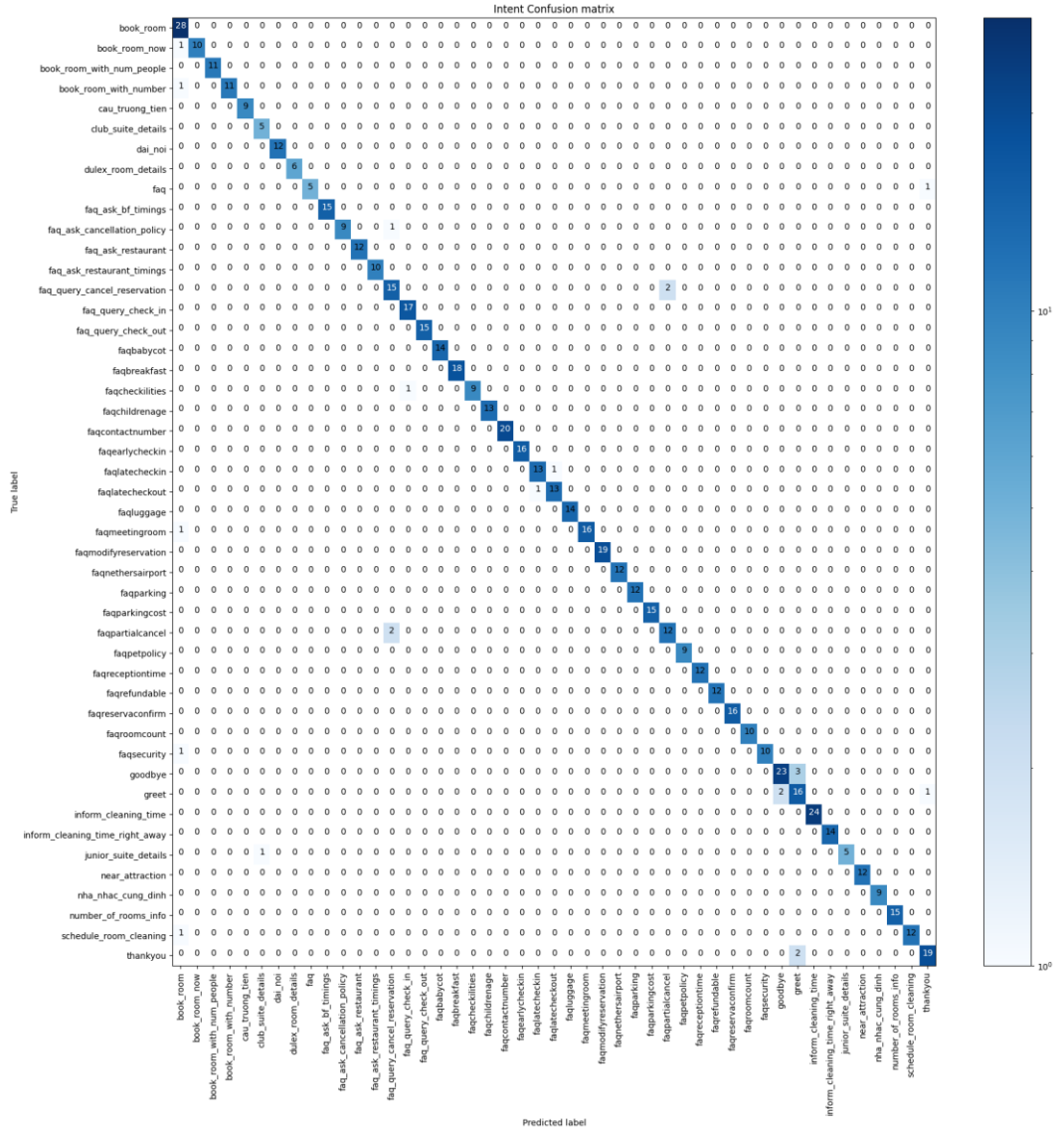
Hình 3. So sánh ba mô hình đã được điều chỉnh

Bảng 2 so sánh chỉ số đánh giá với bài toán phân loại ý định trong chatbot cho thấy mô hình PhoBERT vượt trội hơn so với FastText và BERT.

Bảng 2. So sánh chỉ số đánh giá phân loại ý định

Model	Accuracy	Precision	Recall	F1-Score
PhoBert	0,964	0,974	0,964	0,969
Bert	0,791	0,845	0,779	0,811
FastText	0,624	0,754	0,595	0,665

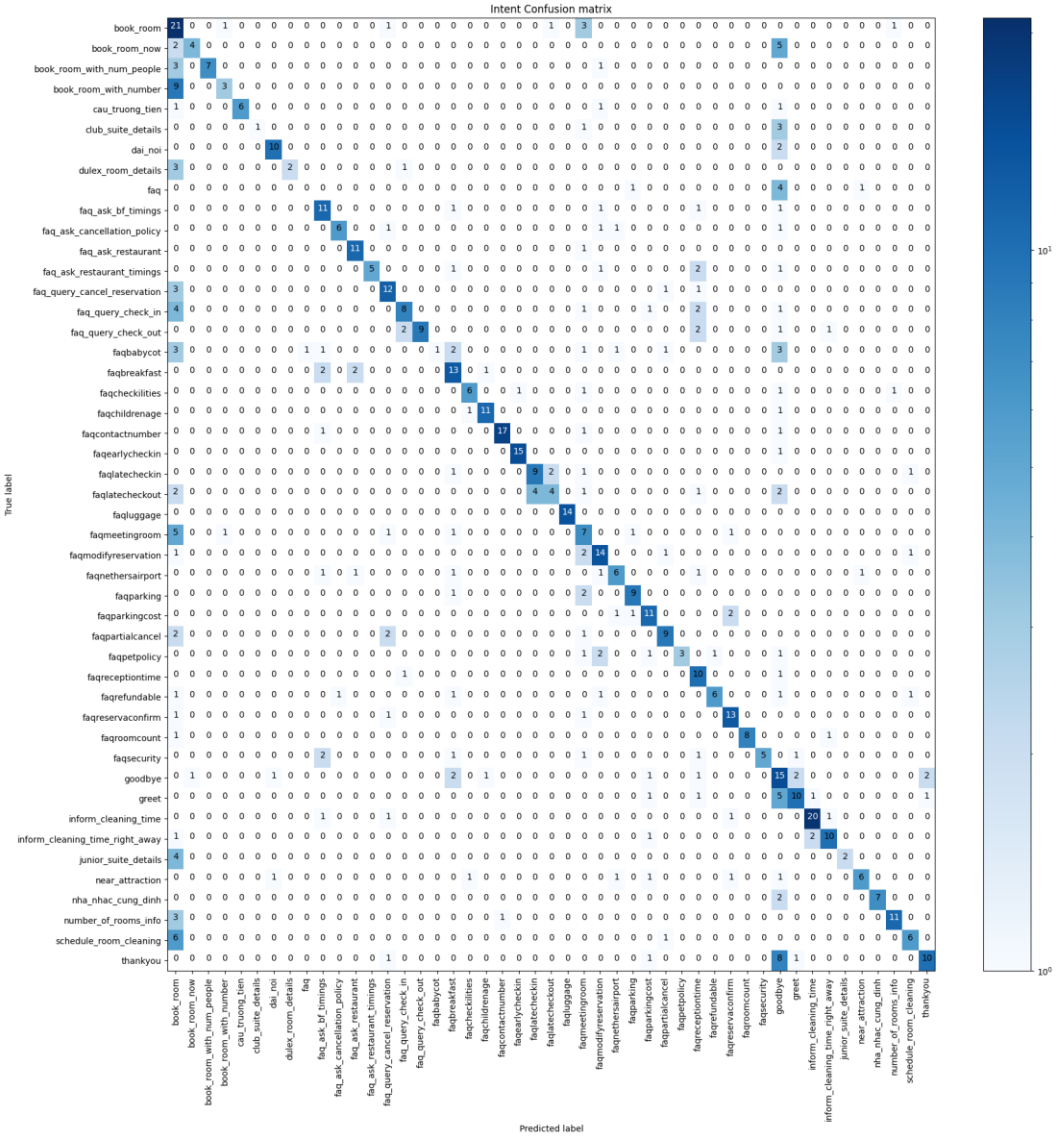
Ma trận nhầm lẫn trên Hình 4 cung cấp thông tin chi tiết cho biết mô hình PhoBERT làm tốt ở đâu, sai nhiều ở đâu, hay giữa những ý định nào xảy ra nhầm lẫn. Trong ma trận nhầm lẫn PhoBERT, ở ý định *"book_room_now"*, mô hình dự đoán đúng 10/11; một trường hợp dự đoán sai, bị nhầm sang ý định *"book_room"*. Tương tự, khi xem xét ở ma trận nhầm lẫn của mô hình FastText trên Hình 4, đối với ý định *"book_room_now"*, mô hình dự đoán đúng 4/11. Bảy trường hợp dự đoán sai gồm hai trường hợp dự đoán nhầm sang ý định *"book_room"* và năm trường hợp nhầm sang ý định *"goodbye"*.



Hình 4. Ma trận nhầm lẫn phân loại ý định theo mô hình PhoBERT

Khi so sánh ma trận nhầm lẫn của PhoBERT trong Hình 4 với FastText trong Hình 5, có thể thấy rằng tổng đường chéo trong ma trận nhầm lẫn của FastText thấp hơn tổng đường chéo của PhoBERT. Điều này cho thấy trong FastText xuất hiện nhiều trường hợp phân loại ý định sai, bị nhầm nhiều ý định hơn so với trong PhoBERT.

Tương tự, ma trận nhầm lẫn của mô hình BERTcho thấy mô hình dự đoán sai nhiều hơn so với PhoBERT sau cải tiến. Như vậy, so sánh ba mô hình FastText, BERT và PhoBERT cải tiến, chúng tôi thấy rằng PhoBERT hoạt động tốt nhất.



Hình 5. Ma trận nhầm lẫn phân loại ý định theo mô hình FastText

5 Kết luận

Chúng tôi đã thành công trong việc mô hình hoá bài toán hiểu tiếng Việt của chatbot trong hỗ trợ thông tin khách sạn. Chúng tôi đã huấn luyện lại mô hình cải tiến PhoBert và phân tích, đánh giá và so sánh với mô hình BERT và FastText. Việc chuẩn bị và chuẩn hóa dữ liệu làm đầu vào cho quá trình điều chỉnh mô hình cũng được mô tả chi tiết; các chỉ số đánh giá được xem xét áp dụng để đánh giá hiệu quả của mô hình. Kết quả là mô hình PhoBERT sau khi thực hiện điều chỉnh cho kết quả tốt nhất trong việc phân loại ý định. Chỉ số đánh giá cho thấy mô hình PhoBERT hoạt động tốt hơn nhiều so với BERT (điểm F1 cao hơn 15,8%) và FastText (điểm F1 là 30,4%). Ma trận nhầm lẫn cũng được chúng tôi sử dụng để đánh giá hiệu quả của ba mô hình khi kích cỡ dữ liệu không thay đổi.

Tài liệu tham khảo

1. Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, Article 634, 7059–7069.
2. B. Athiwaratkun, A. G. Wilson, and A. Anandkumar (2018), Probabilistic fasttext for multi-sense word embeddings, arXiv preprint arXiv:1806.02901, 2018
3. Dat Quoc Nguyen and Anh Tuan Nguyen (2020), PhoBERT: Pre-trained language models for Vietnamese, In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1037–1042, Online. Association for Computational Linguistics.
4. J. Bozic, O. A. Tazl and F. Wotawa, Chatbot Testing Using AI Planning, 2019, IEEE International Conference On Artificial Intelligence Testing (AITest), pp. 37-44, doi: 10.1109/AITest.2019.00-10.
5. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805, 2018.
6. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: a robustly optimized bert pretraining approach. arXiv: Computation and language
7. Nguyễn Ngọc Minh, Nguyễn Quang Huy (2020), Ứng dạng công nghệ chatbot vào du lịch thông minh tại An Giang, Tạp chí công thương, Số 12, Tr 356-361
8. Nguyễn Thanh Thủy (2018), Ứng dụng thuật toán học có giám sát multi-class svm trong xây dựng hệ thống chatbot hỏi đáp tiếng Việt, Kỷ yếu hội thảo khoa học quốc gia 2018 CNTT và ứng dụng trong các lĩnh vực, Tr 177-184
9. N. Le, T. Le, S. Truong and H. Le, Building Filters for Vietnamese Chatbot Responses (2020), 2020 RIVF International Conference on Computing and Communication Technologies (RIVF), 2020, pp. 1-6, doi: 10.1109/RIVF48685.2020.9140770.
10. Oanh Thi Tran, Tho Chi Luong (2020), Understanding what the users say in chatbots: A case study for the Vietnamese language, Engineering Applications of Artificial Intelligence, Volume 87, 2020, 103322, ISSN 0952-1976
11. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. Proceedings of the 2018 Conference of the North American

Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (pp. 2227–2237).

12. P. Gambhir (2019), In: Proceeding of artificial intelligence and speech technology, Indira Gandhi Delhi Technical University for Women, Delhi
13. P. Suanpang and P. Jamjuntr (2021), A chatbot prototype by deep learning supporting tourism, Psychology and Education, 58(4), 1902-1911
14. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog, https://openai.com/blog/better-language-models/
15. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.
16. S. Saradha, M. Sathish, Robin Rathaya, B. Mariyappan and B. S. Akash (2019), Travel Assistant Chatbot System, International Journal of Research in Engineering, Science and Management, Volume-2, Issue-2, February-2019
17. Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson (2018), VnCoreNLP: A Vietnamese Natural Language Processing Toolkit, In Proceedings of NAACL: Demonstrations, pages 56–60
18. T. Nguyen and M. Shcherbakov (2021), Enhancing rasa nlu model for vietnamese chatbot, International Journal of Open Information Technologies, vol. 9, no. 1, pp. 31–36, 2021.
19. Trieu Hai Nguyen, Thi-Kim-Ngoan Pham,Thi-Hong-Minh Bui,Thanh- Quynh- Chau Nguyen (2022), Clustering Vietnamese conversations from Facebook page to build training dataset for chatbot, Jordanian Journal of Computers and Information Technology (JJCIT) ,Volume 08, Number 01, pp. 1 - 17, March 2022, doi: 10.5455/jjcit.71-1632557439.
20. Trisha K R, Ebina S, Sahana Akshadha J, Mrs. T. Subashini (2022), Chatbot Application for Tourism Using Deep Learning, In International Journal for Research in Applied Science and Engineering Technology (Vol. 10, Issue 6, pp. 2661–2663)
21. Vu Anh (2019), UnderTheSea, PyPI. [Online]. Available: https://pypi.org/project/underthesea/. [Accessed: 25-Sep-2021]
22. Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: generalized autoregressive pretraining for language understanding. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, Article 517, 5753–5763.