Hotel Recommendation System with Vietnamese Reviews Based on Deep Learning

Nguyễn Hoàng Minh[1,2], Tạ Nhật Minh[1,2], Nguyễn Thiện Thuật[1,2] and Nguyễn Minh Tiến[1,2]

[1]University of Information Technology, Ho Chi Minh, Vietnam.
[2]Vietnam National University, Ho Chi Minh City, Vietnam.

Contributing authors: 20521609@uit.edu.vn; 20521614@uit.edu.vn; 20521998@uit.edu.vn; 20522010@uit.edu.vn;

Tóm tắt nội dung

Abstract. Du lịch là hoạt động thiết yếu của con người nhằm đáp ứng nhu cầu tham quan, nghỉ dưỡng, tìm hiểu và khám phá. Trong đó, chỗ ở, cụ thể là khách sạn là một trong những yếu tố quan trọng mà con người phải nghĩ đến khi đi du lịch. Chính vì vậy, để nâng cao trải nghiệm của chuyến đi, chúng tôi đã nghiên cứu và đề xuất một giải pháp xây dựng mô hình khuyến nghị từ ngôn ngữ tiếng Việt và dữ liệu người dùng nhằm hỗ trợ họ trong việc chọn khách sạn mong muốn. Dữ liệu của chúng tôi được thu thập trên 2 trang web traveloka và ivivu về các khách sạn ở Việt Nam bao gồm lịch sử feedback người dùng như: comment, rating, tên user, tên khách sạn,... Sau đó được chúng tôi tiền xử lý và gán nhãn với độ đồng thuận lần lượt là: Service: 0.89, Infrastructure: 0.84, Sanitary: 0.83, Location: 0.89, Attitude: 0.83. Mô hình khuyến nghị (recommend model) xây dựng dựa trên Collaborative Filtering Based và deep-learning. Ngoài ra, chúng tôi đề xuất thêm vec-tơ context từ comment tiếng Việt của du khách để tham gia vào quá trình đề xuất. Context model (NLP model) được xây dựng dựa trên deep learning (như một mô hình thứ cấp) nhằm trích xuất các topic và sentiment có trong comment của du khách một cách tốt nhất. Kết quả của mô hình đề xuất trên bộ dữ liệu của chúng tôi thông qua độ đo MSE là 0.027 tốt hơn nhiều so với mô hình không có context được xây dựng với tham số tương tự với độ đo MSE là 0.061. Với NLP model, mô hình được xây dựng sử dụng PhoBERT embedding cho độ chính xác với topic là 81% và sentiment là 82%, con số này với FastText lần lượt là 82% và 81%. Nghiên cứu được thực hiện cho thấy cách tiếp cận của chúng tôi giúp các mô hình khuyến nghị dự đoán tốt hơn và có thể xử dụng để phát triển

thêm trong tương lai. Chúng tôi tin rằng với ý tưởng này sẽ giới thiệu một Recommend System có thể được sử dụng để giải quyết những hạn chế còn tồn tại và hơn thế nữa là mở rộng khả năng ứng dụng của nó.

Keywords: Recommend system, Vietnamese review, PhoBERT, FastText, sentiment analysis, deep learning, LSTM

1 Introduction

Hiện nay, với sự phát triển của hệ thống vận tải giúp cho việc di chuyển dễ dàng hơn, điều này đã và đang thúc đẩy sự phát triển của ngành du lịch. Sự phát triển này có đóng góp lớn cho nền kinh tế phát triển. Du lịch gồm có năm phần chính: địa điểm tham quan, dịch vụ và cơ sở vật chất, giao thông vận tải, thông tin và vị trí, và khách du lịch. Những khía cạnh này được nhiều nhà cung cấp dịch vụ khai thác và áp dụng vào kinh doanh. Tuy nhiên, số lượng nhà cung cấp dịch vụ ngày càng lớn dẫn đến nhiều khó khăn để lựa chọn cho một chuyến du lịch phù hợp với kế hoạch đã đặt ra. Du lịch điện tử (e-travel) đã trở thành một xu hướng mới nổi trên đa nền tảng, cung cấp những tư vấn liên quan đến các tour du lịch, phân tích quan tâm của người dùng để có thể đề xuất các điểm du lịch, gói du lịch, giảm chi tiêu hoặc lịch trình hợp lý và các điểm ưa thích (POI) phù hợp với sở thích của người dùng được đề xuất thông qua du lịch điện tử. Để làm cho mọi kế hoạch dễ dàng thuận tiện và làm giảm độ phức tạp vào việc xác định kế hoạch du lịch sao cho phù hợp, các hệ thống đề xuất (RS) đã được phát triển. Các hệ khuyến nghị (RS) ngày nay đã có thể tích hợp cá nhân hóa du lịch. Nhiều nhà nghiên cứu đã khám phá các khía cạnh khác nhau của các đề xuất dựa trên du lịch; ví dụ: đề xuất khách sạn dựa trên blog của người dùng; giới thiệu nhà hàng dựa trên sở thích của khách hàng; đề xuất POI dựa trên thông tin thời tiết; khuyến nghị gói du lịch tùy chỉnh; dựa trên cuộc trò chuyện để khuyến nghị cho các nhóm du lịch. Các RS như vậy có thể được phát triển cho các ứng dụng dựa trên web, ngoài ra còn có trên thiết bị di động. Trong du lịch điện tử, việc lựa chọn khách sạn được xác định bằng cách kiểm tra các khía cạnh khác nhau có liên quan đến chức năng của chúng và ngữ cảnh liên quan. Các tiêu chí lựa chọn như vậy rất hữu ích để lấy được kỳ vọng của người dùng để có thể đề xuất đúng với mong muốn của họ. Một RS là một phân lớp của hệ thống lọc thông tin dự đoán sở thích của một cá nhân và đề xuất ra một danh sách các phương án phù hợp; các đề xuất sẽ được tạo ra và đưa đến người dùng những sự lựa chọn cụ thể phù hợp với sở thích của họ. Ở bài báo cáo này, chúng tôi đề xuất một hệ thống khuyến nghị về khách sạn dựa trên lịch sử người dùng và context được trích xuất từ comment tiếng Việt như một công cụ giúp máy có thể hiểu được kì vọng của du khách, giúp cho hệ thống hoạt động tốt hơn.

2 Related work

A Comprehensive Survey on Travel Recommender Systems[1]

Ở phần Collaborative Filtering Based Hotel Recocommendations, đã đề cập tới nhiều phương pháp tiếp cận như neuro-fuzzy inference system (ANFIS), clustering, phân tích thành phần chính (PCA), expectation maximization (EM),... cho đề xuất khách sạn. Họ đã cố gắng khắc phục vấn đề đa cộng tuyến của ANFIS bằng cách sử dụng PCA; các quy tắc mờ được trích xuất đã được sử dụng để dự đoán các xếp hạng chưa biết và để tiết lộ mức độ ưa thích của người dùng đối với các feature của item.

Văn bản ngôn ngữ tự nhiên trong các comment được phân tích bằng cách sử dụng phân bổ latent Dirichlet allocation (LDA) và các chủ đề được trích xuất tự động; đối với mỗi chủ đề được trích xuất thì họ sẽ lấy một giá trị tình cảm. Sử dụng item-based CF trong phương pháp được đề xuất, các điểm tương đồng của các khách sạn sẽ được tính toán và xếp hạng cho khách sạn được nhắm mục tiêu đã được dự đoán. Phương pháp được đánh giá trên tập dữ liệu chứa 2256 đánh giá từ TripAdvisor.com với mật độ dữ liệu xấp xỉ 3,57%. Các bước tiền xử lý bao gồm tokenize, stemming và loại bỏ stop word bằng công cụ Stanford CoreNLP và nhiều topic khác nhau đã được chọn cho thử nghiệm; trung bình 0,67 (MAE) và 0,87 lỗi bình phương gốc (RMSE) đã được tìm thấy cho phương pháp đề xuất.

Restaurant recommender system based on sentiment analysis[2]

Trong bài báo này, một hệ thống khuyến nghị nhận biết ngữ cảnh được đề xuất rằng cách trích xuất sở thích ăn uống của các cá nhân từ nhận xét của họ và đề xuất các nhà hàng phù hợp với những sở thích này. Trang web TripAdvisor đã được sử dụng và nhận xét từ 100 người dùng khác nhau đã được thu thập trong 9 tháng đầu năm 2018. Độ chính xác, recall, f-đo measure của hệ thống được đo lường theo ba bản top1, top3 và top5. Kết quả chỉ ra rằng hệ thống được đề xuất có thể cung cấp các đề xuất với độ chính xác 92,8%, mang lại cho người dùng với độ chính xác cao.

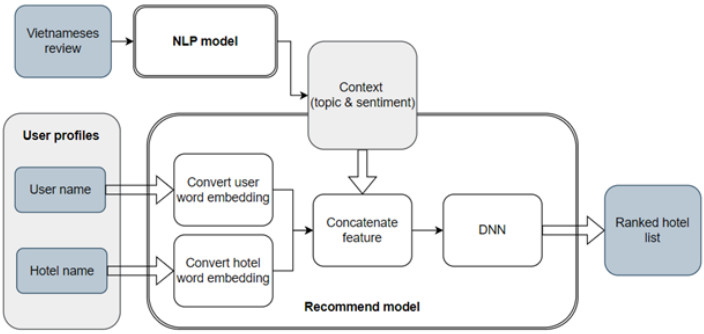
Learner comments-based Recommendation system[3]

Đề tài xây dựng một mô hình đề xuất dựa trên các bài đánh giá bằng văn bản do người dùng viết trực tuyến trên các trang web các khóa học online. Mô hình được xây dựng dựa trên hai phương pháp khuyến nghị (recommend) và phân tích tình cảm (sentiment analysis). Nó đã được training và thử nghiệm dựa trên nhận xét và đánh giá của người học trực tuyến về việc học qua video của ba khóa học trên nền tảng giáo dục Coursera để tạo ra một hệ thống khuyến nghị học tập qua video, xác nhận tính chính xác và hiệu quả của mô hình được đề xuất.

3 Proposed aproach

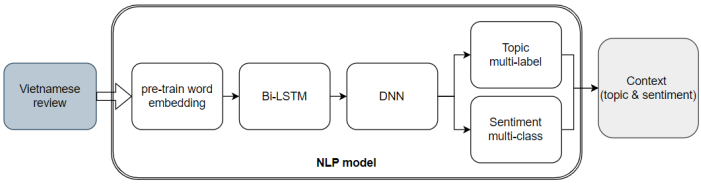
Cách tiếp cận của chúng tôi là sự kết hợp giữa hệ thống đề xuất (recommend systems) dựa trên "Collaborative Filtering Based Hotel" với mạng nơ-ron nhân

tạo (deep learning). Đồng thời, các comment feedback của người dùng cũng được sử dụng thêm làm context cho hệ thống (hình 1).



Hình 1: Sơ đồ tổng thể của mô hình khuyến nghị

Hình 2 là sơ đồ mô hình xử lý ngôn ngữ nhằm trích xuất các đặc trưng từ comment của khách hàng sau khi review khách sạn.



Hình 2: Sơ đồ tổng thể của mô hình NLP

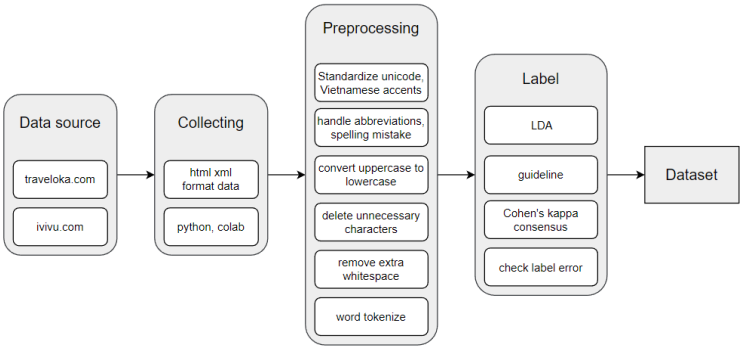
Hệ thống đề xuất có quy trình chạy theo 2 bước như sau:

- Mô hình xử lý ngôn ngữ sẽ học tất cả các comment của khách hàng nhằm dự đoán tốt nhất các topic có xuất hiện trong câu hay không và mức độ thể hiện sự hài lòng hay không hài lòng của khách hàng thông qua comment đó để làm context cho hệ thống đề xuất.
- Mô hình recommend sẽ cố gắng dự đoán rating từ việc học các profile của người dùng nhằm đưa những khách hàng có cùng sở thích, đặc điểm lại gần nhau hơn trong không gian vectơ thông qua embedding (giống cách tiếp cận của Collaborative Filtering Based), với input đầu vào là lịch sử profiles của người dùng và comment về review khách sạn gần nhất của khách hàng.

Output context bao gồm các topic được đề cập nhiều trong các comment của khách hàng và cảm xúc chung của toàn câu.

Output của hệ thống recommend là một danh sách các khách sạn sắp xếp theo thứ tự giảm dần dựa trên rating được mô hình đề xuất.

4 Dataset creation



Hình 3: Quy trình thu thập và tiền xử lý dữ liệu

4.1 Collecting & preprocessing

Quá trình thu thập dữ liệu cho dự án này được thực hiện bằng những phương pháp thu thập trên website và tiến hành ứng dụng để có được bình luận từ những trang nổi tiếng trong việc đặt phòng khách sạn ở Việt Nam. Để có được bộ dữ liệu, chúng tôi đã thực hiện thăm dò dữ liệu của nhiều website khác nhau nhằm tìm ra nơi cung cấp những bình luận phù hợp và những thông tin khác có ý nghĩa cho việc phát triển mô hình.

Sau khi có được bộ dữ liệu thô chúng tôi đã tiền xử lý. Các comment được người dùng viết trên nhiều thiết bị và với những khoảng thời gian khác nhau. Điều này dẫn đến việc không đồng bộ về unicode hay dấu câu của các từ trong bộ dự liệu, dẫn đến những từ khi được con người nhìn vào thì giống nhau nhưng khi để cho máy nhận dạng thì có thể khác nhau, khiến cho việc dự đoán mang về kết quả thấp hơn so với những gì có thể đạt được. Vì vậy cần phải chuẩn hóa unicode, dấu trong tiếng Việt. Việc tiếp theo sẽ khó khăn hơn khi ngay cả con người cũng có thể không nhận biết đó chính là chữ viết tắt, không dấu, lỗi chính tả, spam hoặc những từ ngoại quốc. Đây là giai đoạn tốn nhiều thời gian và công sức nhất trong quá trình xử lý dữ liệu. Vì lí do con người còn khó để hiểu đúng về ý nghĩa của từng từ, từng câu nên không thể xử lý hoàn toàn bằng máy mà phải thực hiện xử lý bán tự động và kiểm duyệt lại bộ dữ liệu nhiều lần. Sau giai đoạn khó khăn này, những bước còn lại trở nên dễ dàng hơn khi hoàn toàn có thể để máy xử lý tự động như là đưa các kí tự chữ in hoa về chữ thường, xóa các ký tự không cần thiết (những dấu câu dư thừa, các icon bộc lộ cảm xúc) hay xóa đi các khoảng trắng thừa. Phần tiếp theo sẽ giúp cho mô hình có thể xác định tốt hơn ý nghĩa của từ vựng chính là tách từ (word tokenize). Khi đó một từ được tạo nên từ hai tiếng hoặc hơn sẽ được kết nối lại với nhau để thể hiện rõ đó chỉ là một từ.

Bảng 1: Dữ liệu sau khi được tiền xử lý

User_name	Location	Hotel_name	Rating	Comment
Nguyen V. C.	Phan Thiết	Khu nghỉ dưỡng Pandanus Phan Thiết	10.0	gia_đình đi ok chuyến đi thành_công vui_vẻ
Nguyen T.A.T.	Phú Quốc	Vinpearl Resort & Spa Phú Quốc	8.5	khách_sạn dịch_vụ tốt thủ_tục gọn nhân_viên nhiệt_tình vui_vẻ phòng trễ khách_sạn phòng nghỉ phòng thuê nhân_viên ivivu tư_vấn nhiệt_tình
Dang T.D.	Đà Lạt	Khách Sạn LADALAT	6.0	khách_sạn hơi trung_tâm phòng_ốc tốt buffet món thức_ăn tệ không_ngon nhân_viên tư_vấn tốt

Cuối cùng ta đã có được bộ dữ liệu sạch với 12696 bình luận với 5 thuộc tính. Nhưng để tránh bị nhiễu từ những từ không mang nhiều giá trị về mặt ngữ nghĩa thì cần phải làm thêm một bước xử lý dữ liệu để loại bọ chúng (stop word). Tuy nhiên quá trình này có thể khiến cho việc gán nhãn dữ liệu trở nên khó khăn hơn đôi chút nên chúng tôi đã tiến hành gán nhãn dữ liệu trước khi thực hiện thao tác này.

4.2 Labeling

Sử dụng mô hình LDA[4] (Latent Dirichlet Allocation) để xác định tập hợp các từ, cụm từ có trong bộ dữ liệu, từ đó tiến hành chọn các chủ đề liên quan đến bộ dữ liệu để tiến hành việc gán nhãn. Sau khi tiến hành, chúng tôi tập hợp và phân chia bộ dữ liệu thành bốn chủ đề: $Service(dchv), Infrastucture(csvtcht), Sanitary(vsinh), Location(vtr)$. Mỗi bình luận có thể không thuộc chủ đề nào hoặc thuộc một hay nhiều chủ đề.

Dữ liệu được phân chia thành năm nhóm nhãn, với bốn nhóm nhãn đầu tiên tương ứng với bốn chủ đề đã đặt ra với nhãn '0' tương ứng với việc bình luận không thuộc chủ đề và nhãn '1' tương ứng với việc bình luận thuộc chủ đề đó. Nhóm nhãn thứ 5 là nhãn thể hiện độ hài lòng của người bình luận đối với khách sạn, với các nhãn Unsatisfied '0' thể hiện sự không hài lòng, Normal '1' thể hiện thái độ bình thường, Satisfied '2' thể hiện thái độ hài lòng. Chúng tôi thực hiện gán nhãn cho 2231 bình luận đầu tiên trong bộ dữ liệu. Bộ dữ liệu được gán nhãn bởi 2 annotators, và độ đồng thuận được tính dựa trên công thức Cohen's Kappa[5]:

$$k = \frac{Pr(A) - Pr(e)}{1 - Pr(e)} \tag{1}$$

Bảng 2: Bảng phân loại chủ đề của bộ dữ liệu

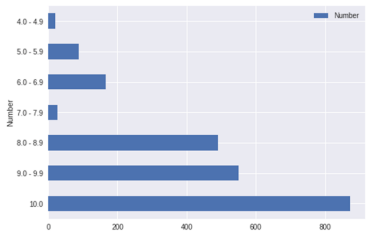
Chủ đề	Nội dung	Keyword
Service	Các bình luận đánh giá về dịch vụ của khách sạn	nhân viên, lễ tân, lịch sự, phục vụ, tư vấn, dịch vụ, hỗ trợ, phản hồi, thức ăn,..
Infrastructure	Các bình luận đánh giá về cơ sở vật chất của khách sạn	phòng, khách sạn, bàn ghế, giường ngủ, phòng tắm, rộng rãi, tiện nghi,..
Sanitary	Các bình luận đánh giá về vấn đề vệ sinh của khách sạn	nhà vệ sinh, sạch sẽ, sạch đẹp, thoáng mát, dọn dẹp, vệ sinh,...
Location	Các bình luận đánh giá về vị trí của khách sạn	view đẹp, trung tâm, vị trí, khung cảnh, biển, núi, sông, hồ, gần với,...

Sau khi tiến hành gán nhãn lần một, kết quả thể hiện ở Bảng 3, chúng tôi nhận thấy độ đồng thuận chưa cao và có thể ảnh hưởng đến kết quả của mô hình. Vì vậy, chúng tôi tiến hành gán nhãn lần hai trên tập các nhãn bị gán sai hoặc có khắc biệt quá lớn giữa 2 annotators.

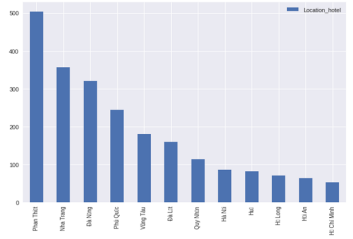
Bảng 3: Kết quả độ đồng thuận của quá trình gán nhãn dữ liệu

	Service	Infrastructure	Sanitary	Location	Attitude
Lần 1	$k = 0.796$	$k = 0.718$	$k = 0.724$	$k = 0.766$	$k = 0.695$
Lần 2	$k = 0.893$	$k = 0.841$	$k = 0.828$	$k = 0.886$	$k = 0.827$

4.3 Dataset result



Hình 4: Biểu đồ đánh giá chất lượng khách sạn



Hình 5: Biểu đồ số lương khách sạn được khảo sát theo địa điểm

Bộ dữ liệu hoàn chỉnh bao gồm 2231 dòng dữ liệu tương ứng với 2231 bình luận đánh giá khách sạn đã được gán nhãn. Trong đó có 368 user, 137 hotel

khác nhau và Rating nằm trong khoảng 4.5 - 10. Với số lượng các đánh trên thang điểm 10 được mô tả trong Hình 4 và số lượng khách sạn được đánh giá tại 12 địa điểm được mô tả trong Hình 5. Bảng 4 thể hiện số lượng bình luận được gán nhãn theo từng chủ đề và thái độ đánh giá của người dùng.

Bảng 4: Bảng thống kê số lượng nhãn theo chủ đề và độ hài lòng

Chủ đề	Số lượng đánh giá hài lòng	Số lượng đánh giá bình thường	Số lượng đánh giá không hài lòng	Số bình luận được gán nhãn/Tổng số bình luận được gán
Service	1237	199	191	1627/2231
Infrastructure	952	200	192	1344/2231
Sanitary	463	75	108	646/2231
Location	348	82	35	465/2231
Không xác định	204	31	6	241/2231

5 Model neural network design

5.1 Backbone

5.1.1 PhoBert Model

PhoBert[6] là một pre-trained được huấn luyện dành riêng cho tiếng Việt. Việc huấn luyện dựa trên kiến trúc và cách tiếp cận giống RoBERTa của Facebook được Facebook giới thiệu giữa năm 2019 (đây là một cái tiến hơn so với BERT trước đây). Tương tự như BERT, PhoBERT cũng có 2 phiên bản là PhoBERT_{base} với 12 transformers block và PhoBERT_{large} với 24 transformers block. PhoBERT được train trên khoảng 20GB dữ liệu bao gồm khoảng 1GB Vietnamese Wikipedia corpus và 19GB còn lại lấy từ Vietnamese news corpus. Đây là một lượng dữ liệu khả ổn để train một mô hình như BERT. PhoBERT sử dụng RDRSegmenter của VnCoreNLP để tách từ cho dữ liệu đầu vào trước khi qua BPE encoder. Như đã nói ở trên, do tiếp cận theo tư tưởng của RoBERTa, PhoBERT chỉ sử dụng task Masked Language Model để train, bỏ đi task Next Sentence Prediction.

5.1.2 FastText Model

Trong fastText[7], mỗi từ trung tâm được biểu diễn như một tập hợp của các từ con được trích xuất n-grams. Với một từ w, ta ghi tập hợp của tất cả các từ con của nó với chiều dài từ 3 đến 6 và các từ con đặc biệt là G_w . Do đó, từ điển này là tập hợp các từ con của tất cả các từ. Giả sử vector của từ con g trong từ điển này là z_g . Thì vector từ trung tâm u_w cho từ w trong mô hình

skip-gram có thể biểu diễn là:

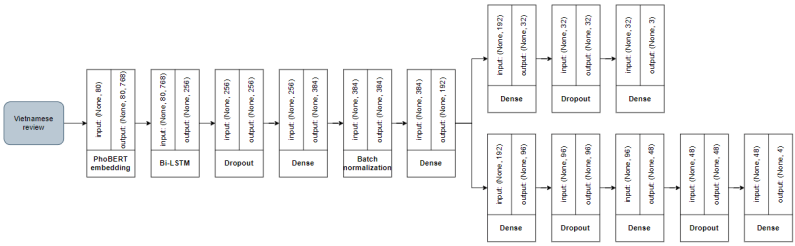
$$u_w = \sum_{g \in g_w} z_g \tag{2}$$

Phần còn lại của tiến trình xử lý trong fastText đồng nhất với mô hình skip-gram[8]. Từ điển của fastText lớn hơn dẫn tới nhiều tham số mô hình hơn. Hơn nữa, vector của một từ đòi hỏi tính tổng của tất cả vector từ con dẫn tới độ phức tạp tính toán cao hơn. Tuy nhiên, ta có thể thu được các vector tốt hơn cho nhiều từ phức hợp ít thông dụng, thậm chí cho cả các từ không hiện diện trong từ điển này nhờ tham chiếu tới các từ khác có cấu trúc tương tự.

5.1.3 LSTM Model

Mạng trí nhớ ngắn hạn định hướng dài hạn còn được viết tắt là LSTM[9] là một kiến trúc đặc biệt của RNN có khả năng học được sự phục thuộc trong dài hạn (*long − termdependencies*) được giới thiệu bởi Hochreiter & Schmidhuber (1997). Kiến trúc này đã được phổ biến và sử dụng rộng rãi cho tới ngày nay. LSTM đã tỏ ra khắc phục được rất nhiều những hạn chế của RNN trước đây về triệt tiêu đạo hàm. Tuy nhiên cấu trúc của chúng có phần phức tạp hơn mặc dù vẫn dữ được tư tưởng chính của RNN là sự sao chép các kiến trúc theo dạng chuỗi

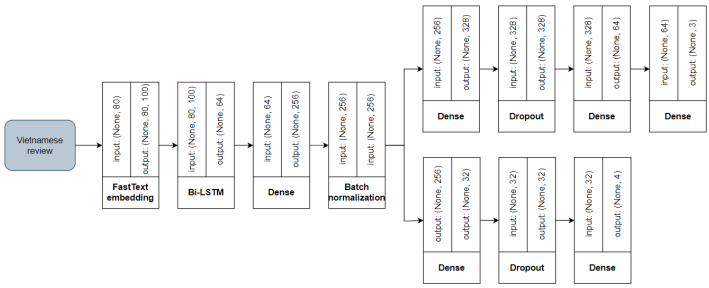
5.2 Topic & sentiment classification task



Hình 6: Kiến trúc mạng deep-learning sử dụng PhoBERT

Về phần chuẩn bị dữ liệu cho việc trainning mô hình, chúng tôi đã tokenize riêng để phù hợp với từng mô hình pre-train word emmbedding. Sau đó tiến hành khảo sát độ dài câu phù hợp và padding để đưa về độ dài câu chuẩn cho toàn bộ dataset, trong bài báo cáo này chúng tôi chọn 80. Dữ liệu được chia thành tập train và tập test với tỉ lệ 4:1.

Mô hình mạng được thiết kế sử dụng 5 thành phần chính embedding, LTSM, Dropout, Dense, Batch normalization với mục tiêu giúp máy tính có thể học và phân loại chính xác được 2 đầu ra là 4 topic: Service, Infrastructure, Sanitary, Location và 3 trạng thái của Attitude: 0: không hài lòng, 1: trung tính, 2: hài lòng ứng với mỗi comment.



Hình 7: Kiến trúc mạng deep-learning sử dụng FastText

Bảng 5: Mẫu ví dụ

Sample	Service	Infrastructure	Sanitary	Location	Attitude
nhân_viên nhiệt_tình khách_sạn sạch_sẽ	1	0	1	0	2

Các lớp dropout, batch normalization được thêm vào giữa các layer với mong muốn giảm bỏ các vấn đề mô hình bị overfiting khi trainning dữ liệu với số lần lặp lớn. Lớp Bi-LSTM được thêm vào với kỳ vọng là dữ liệu đầu ra của các embedding có thể tìm được mối liên quan giữa các từ xung quanh một cách tuần tự và LSTm được chọn nhằm có thể nhớ được các phụ thược của các từ ở xa vì dữ liệu padding của dataset chúng tôi tương đối dài (80 kí tự).

5.3 Recommendation task

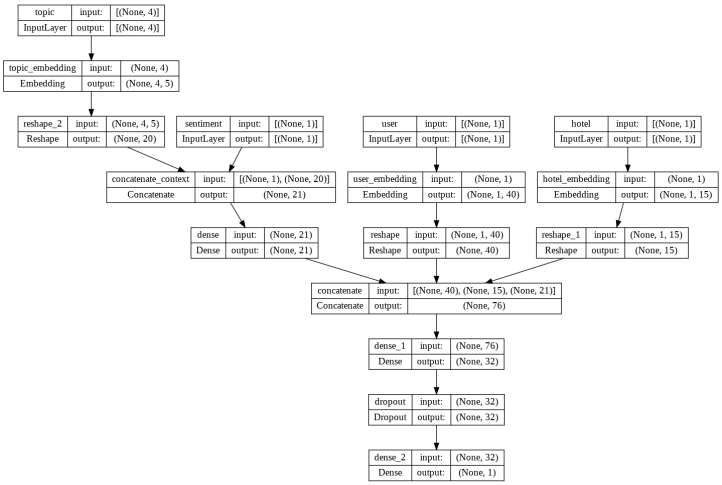
Dữ liệu cho việc trainning mô hình được chúng tôi chia thành 2 nhóm và được chia thành 2 tập train và test với tỉ lệ 4:1 bao gồm:

• Dữ liệu lịch sử người dùng: bao gồm feedback của người dùng về khách sạn của họ đã trải nghiệm

• Dữ liệu context trích xuất được từ comment: bao gồm 4 topic và 3 class của thuộc tính Attitude

Mô hình được thiết kế với các thành phần tương tự như thành phần của NLP model phía trên. Cell embeding được sử dụng như một colaborative filtering nhằm đưa những khác hàng có lịch sử và đặt điểm gần nhau hơn trong không gian với mục đích có thể gợi ý được các khách sạn mà những người có điểm tương đồng chưa trải nghiệm.

Context là một giải pháp của chúng tôi đề xuất và được thêm vào nhằm xác định được ngữ cảnh và trạng thái của người dùng sau khi trải nghiệm khách sạn với mong muốn là mô hình có thể dự đoán tốt hơn. Context này bao gồm 2 yếu tố là topic và cảm xúc của họ được dự đoán từ commment thông qua NLP model. Topic quan tâm tới khía cạnh chủ đề hiện tại và sentiment quan tâm đến khía cạnh cảm xúc chung của họ, sau đó được concatenate lại là xem như là một context hỗ trợ cho recommend model.



Hình 8: Kiến trúc mạng deeplearning recommendation model

Mô hình sẽ cố gắn dự đoán kết quả rating của người dùng về các khách sạn mới mà họ chưa đi thông qua context của họ. Rating này đã được chúng tôi Normalize vào trong khoảng [0,1] nhằm giúp mô hình dễ dàng hơn trong việc tranning bằng công thức dưới đây:

$$\frac{x - min_{rating}}{max_{rating} - min_{rating}} \tag{3}$$

Bảng 6: Mẫu ví dụ

User	Hotel	Service	Infrast	Sanitary	Location	Attitude	Rating
Thao T.	Khách sạn Thanh Bình 2	1	0	1	0	2	9.7

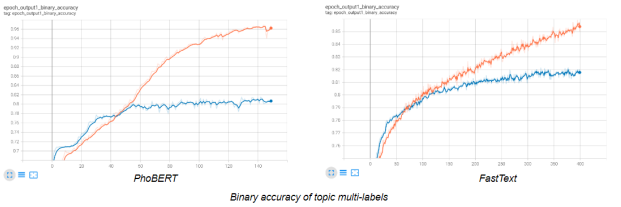
6 Experiments & disscussion

6.1 Topic & sentiment classification task

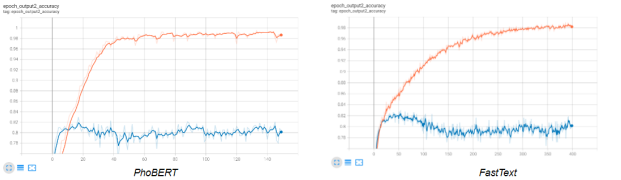
Các độ đo đánh giá của chúng tôi cho bài toán này bao gồm binary accuracy cho bài toán topic multi-labels và accuracy cho bài toán sentiment multi-class. Kết quả của các mô hình giao động trong khoảng 80% đến 82%.

Lỗi dự đoán từ mô hình

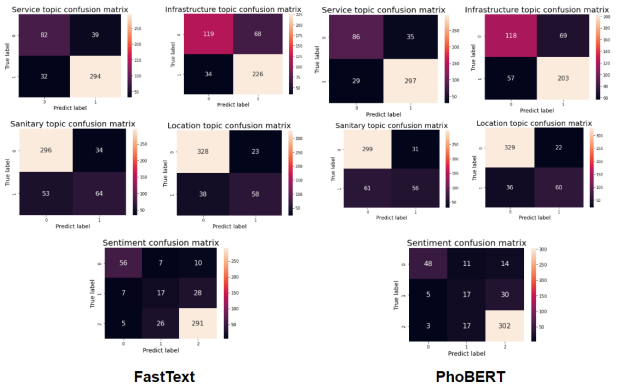
Sau khi xây dựng mô hình chúng tôi nhận thấy được những lỗi mà mô hình dự đoán chưa chính xác:



Binary accuracy of topic multi-labels



Accuracy of sentiment multi-class



Confusion matrix

Service					
	Precision	Recall	F1-score	Accuracy	Macro average
0	0.75	0.71	0.73	0.86	0.82
1	0.89	0.91	0.90		
Infrastructure					
0	0.67	0.63	0.65	0.72	0.71
1	0.75	0.78	0.76		
Sanitary					
0	0.83	0.91	0.87	0.79	0.71
1	0.64	0.48	0.55		
Location					
0	0.90	0.94	0.92	0.87	0.80
1	0.73	0.62	0.67		
Sentiment					
0	0.86	0.66	0.74	0.82	0.67
1	0.38	0.33	0.35		
2	0.87	0.94	0.90		

PhoBERT

Service					
	Precision	Recall	F1-score	Accuracy	Macro average
0	0.72	0.68	0.70	0.84	0.80
1	0.88	0.90	0.89		
Infrastructure					
0	0.78	0.64	0.70	0.77	0.76
1	0.77	0.87	0.82		
Sanitary					
0	0.85	0.90	0.87	0.81	0.73
1	0.65	0.55	0.60		
Location					
0	0.90	0.93	0.91	0.86	0.79
1	0.72	0.60	0.66		
Sentiment					
0	0.82	0.77	0.79	0.81	0.67
1	0.34	0.33	0.33		
2	0.88	0.90	0.89		

FastText

Classification report

- Đối với mô hình PhoBert: Khó nhận dạng attitude trung tính. Không nhận dạng được từ chung. Không phát hiện được topic của bình luận đối với Sanitary và Location.

- Đối với mô hình FastText: Trung tính nhận nhầm qua hài lòng nhiều. Không phát hiện được topic của bình luận đối với Sanitary và Location.

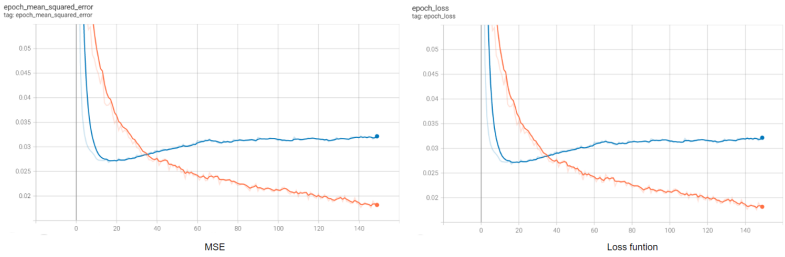
Sentence	khi nhận phòng tại đây chúng_tôi phải chuyển phòng vài lần phòng thị có mùi thuốc_lá do không_vệ_sinh kỹ phòng thì bị hôi cống thoát nước nói_chung là không_hài_lòng khi nghĩ tại đây				
Label	0	1	1	0	0
PhoBert	1	0	0	1	0
FastText	1	1	0	0	0
Sentence	nhân_viên rất thân_thiện thái_độ phục_vụ chu_đáo nhà_hàng nấu_ăn ngon và sạch_sẽ khung_cảnh yên_tĩnh phù_hợp với nghỉ_dưỡng				
Label	1	0	1	1	2
PhoBert	1	0	0	1	2
FastText	1	1	1	0	2
Sentence	lúc tôi gọi món ăn ở nhà_hàng thì các bạn không_hiểu ý lắm buổi tôi tôi có mượn 2 cái tô để sử_dụng nhưng chờ hơn 1 tiếng cũng chưa thấy tôi phải gọi nhắc chỉ có 2 vấn_đề trên còn phòng đẹp nhìn_chung phục_vụ cũng ok				
Label	1	1	0	0	1
PhoBert	1	0	0	0	1
FastText	1	0	0	0	2

Hình 9: Dự đoán mẫu ngẫu nhiên

6.2 Recommendation task

Để đánh giá kết quả dự đoán của mô hình chúng tôi sử dụng độ đo MAE và MSE. Trong đó MSE dùng làm độ đo chính của mô hình có công thức như sau:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \tag{4}$$



Hình 10: Kết quả trainning của Recommend model

Từ kết quả mô hình trên, model có xu hướng giảm mạnh ở khoảng 30 epoch đầu tiên và đạt hiểu quá tốt nhất với MSE là 0.0268. Sau đó mặc dù hàm lỗi ở tập train có xu hướng giảm tiếp, tuy nhiên trên tập test đã bắt đầu bão hòa trong 0,03 đến 0,32.

Chúng tôi đã thực hiện kiểm nghiệm mức độ hiệu quả của mô hình có context bằng cách xây dựng một mô hình recommend khác với kiến trúc và siêu tham số tương tự như mô hình đề xuất. Kết quả tốt nhất của mô hình không có context có MSE là 0.0607, thấp hơn rất nhiều so với mô hình được chúng tôi đề xuất.

Demo kết quả dự đoán của mô hình

UserId	HotelId	Service	Infrastructure	Sanitary	Location	Attitude
Nguyen T. H.	Khu nghỉ dưỡng Pandanus Phan Thiết	1	1	0	0	2
Nguyen T. H.	Victoria Phan Thiết Beach Resort & Spa	1	0	1	1	2
Nguyen T. H.	Khách sạn Hanoi Sunshine	1	1	1	0	2
Nguyen T. H.	Khách sạn DDA Quận 1	0	0	0	1	0
Nguyen T. H.	Khu nghỉ dưỡng Crown Retreat Quy Nhơn	1	1	0	0	2
Nguyen T. H.	Khách sạn Sea Links Beach Phan Thiết	0	0	0	0	1
Nguyen T. H.	Victoria Phan Thiết Beach Resort & Spa	1	1	1	0	2
Nguyen T. H.	Khu nghỉ dưỡng Ana Mandara Villas Dalat Resort...	1	1	0	0	2
Nguyen T. H.	Khu nghỉ dưỡng Pandanus Phan Thiết	1	1	0	0	2
Nguyen T. H.	Khu nghỉ dưỡng Fusion Phú Quốc	0	1	0	0	2
Nguyen T. H.	Khách sạn Mường Thanh Luxury Khánh Hòa	1	1	0	0	1
Nguyen T. H.	Khách sạn Hanoi Sunshine	1	1	1	0	2
Nguyen T. H.	Victoria Phan Thiết Beach Resort & Spa	1	1	1	0	2

Hình 11: Lịch sử khách sạn khách hàng đã bình luận

Bảng 7: 10 khách sạn được mô hình đề xuất

Top 10 hotel recommendations
InterContinental Đà Nẵng Sun Peninsula Resort
Khách sạn A La Carte Đà Nẵng
Vinpearl Discovery Wonderworld Phú Quốc
Khu nghỉ dưỡng Premier Village Phu Quoc Managed By Accor
Khu nghỉ dưỡng Melia Đà Nẵng Beach
Meliã Vinpearl Đà Nẵng Riverfront
Vinpearl Resort Nha Trang - Hon Tre Island
Khu nghỉ dưỡng Amiana Nha Trang
Khu nghỉ dưỡng Six Senses Côn Đảo
Khách sạn Mường Thanh Luxury Quảng Ninh

7 Conclusion

Từ kết quả nghiên cứu của đề tài, chúng tôi đã chứng minh được cách tiếp cận theo context được trích xuất từ comment feedback của người dùng cho kết quả tốt hơn so với mô hình không sử dụng. Đề xuất một giải pháp tiếp cận xây dựng mô hình recommend system từ lịch sử feedback của người dùng trên các trang web du lịch điện tử. Đồng thời, trong quá trình nghiên cứu chúng tôi cũng đóng góp bộ dữ liệu đã được tiền xử lý bao gồm các thông tin như tên user, tên hotel, comment,... Và một phần nhỏ đã được gán nhãn với độ đồng thuận tương đối cao trên 80%

Tuy nhiên, đề tài cũng còn gặp nhiều hạn chế khiến cho chất lượng bộ dữ liệu cũng như phương pháp luận chưa được chặt chẽ. Việc lỗi trong quá trình gán nhãn, sự mất cân bằng về các nhãn trong bộ dữ liệu, việc chọn topic và xây dựng công thức tính toán sentiment cho comment mà trong câu có nhiều cảm xúc với nhiều topic khác nhau được đề cập là những lỗ hổng trong đề tài chưa thực hiện triệt để. Hạn chế về thời gian thực hiện đề tài cũng như giới hạn về kiến thức cũng là một trong những lý do khiến cho bộ dữ liệu độ chính xác của mô hình chưa thể đạt kết quả tốt hơn

Trong tương lai, đề tài có thể được phát triển thêm để hoàn thiện tính chặt chẽ hóa của phương pháp luận như giải quyết các vấn đề mà đề tài này đã gặp nhiều khó khăn gặp phải, thực hiên gán nhãn trên toàn bộ tập dữ liệu đã được tiền xử lý, xây dựng mô hình end to end từ 2 bài toán đã đề xuất, xây dựng web demo kết quả nghiên cứu của đề tài.

Authorship contribution statement

Nguyễn Hoàng Minh: Xây dựng phương pháp, lên cấu trúc báo cáo, xây dựng mô hình, đánh giá báo cáo

Nguyễn Thiện Thuật: Thu thập, tiền xử lý dữ liệu, tạo bộ liệu, viết và chỉnh sửa báo cáo bản gốc

Tạ Nhật Minh: Xây dựng và đánh giá quy trình gán nhãn, tạo bộ dữ liệu, trình bày slide báo cáo

Nguyễn Minh Tiến: Lý thuyết, khái niệm hóa, hỗ trợ

Tài liệu

[1] Chaudhari, K., Thakkar, A.: A comprehensive survey on travel recommender systems. Archives of Computational Methods in Engineering **27**(5), 1545–1571 (2020)

[2] Asani, E., Vahdat-Nejad, H., Sadri, J.: Restaurant recommender system based on sentiment analysis. Machine Learning with Applications **6**, 100114 (2021)

[3] Hazar, M.J., Zrigui, M., Maraoui, M.: Learner comments-based recommendation system. Procedia Computer Science **207**, 2000–2012 (2022)

[4] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation (2003)

[5] Cohen, J. (ed.): A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)

[6] N., D.Q., Nguyen, A.T.: PhoBERT: Pre-trained language models for Vietnamese

[7] Athiwaratkun, B., Wilson, A.G., Anandkumar, A.: Probabilistic fasttext for multi-sense word embeddings. arXiv preprint arXiv:1806.02901 (2018)

[8] Lazaridou, A., Pham, N.T., Baroni, M.: Combining language and vision with a multimodal skip-gram model. arXiv preprint arXiv:1501.02598 (2015)

[9] Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R., Schmidhuber, J.: LSTM: A search space odyssey. IEEE (2016)