

**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**KHOA CÔNG NGHỆ THÔNG TIN**

-----□□□□□-----



**ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC**

**ĐỀ TÀI: MÔ HÌNH KẾT HỢP HÀNH VI ĐÁNH GIÁ**

**VÀ BÌNH LUẬN CHO TƯ VẤN KHÁCH SẠN**

**Giảng viên hướng dẫn: PGS.TS TRẦN ĐÌNH QUẾ**

**Sinh viên: VŨ QUANG SƠN**

**Lớp D17HTTT2**

**Mã sinh viên: B17DCCN545**

**Hệ đại học: ĐẠI HỌC CHÍNH QUY**

**Hà Nội 2021**

## LỜI CẢM ƠN

Lời đầu tiên, em xin gửi lời cảm ơn chân thành tới tất cả thầy cô đang giảng dạy trong mái trường Học viện Công nghệ Bru chính Viễn thông đã tận tình truyền đạt những kinh nghiệm và kiến thức quý báu giúp em hoàn thành nhiệm vụ học tập trong suốt khoảng thời gian hơn 4 năm là sinh viên của học viện. Em xin gửi lời biết ơn sâu sắc đến thầy PGS.TS Trần Đình Quế, người đã tận tình hướng dẫn, chỉ bảo, định hướng và nhắc nhở em trong suốt quá trình học tập cũng như hoàn thành đồ án này.

Cho con gửi lời cảm ơn chân thành đến bố mẹ, ông bà, anh chị em đã luôn động viên, ủng hộ, cổ vũ và tạo điều kiện tốt nhất cho con trong suốt những năm tháng ngồi trên ghế nhà trường.

Cuối cùng, cho tôi gửi lời cảm ơn đến những người bạn, người anh, người chị của tôi, những người luôn chia sẻ, động viên, giúp đỡ và ở bên tôi mỗi khi tôi gặp khó khăn nhất!

Em xin chân thành cảm ơn!

*Hà Nội, ngày 10 tháng 12 năm 2021*

*Sinh viên thực hiện*

**Vũ Quang Sơn**

## NHẬN XÉT

**(Của giảng viên phản biện)**

[illegible]

*Hà Nội, 12/2021*  
**Giảng viên phản biện**

## NHẬN XÉT

**(Của giảng viên hướng dẫn)**

This image shows a full page of white paper with horizontal dotted lines. The lines are evenly spaced and run across the width of the page, providing a guide for handwriting practice. There are no margins, text, or other markings on the page.

*Hà Nội, 12/2021*  
**Giảng viên hướng dẫn**

## MỤC LỤC

LỜI CẢM ƠN.....	i
NHẬN XÉT.....	ii
NHẬN XÉT.....	iii
MỤC LỤC.....	iv
DANH MỤC CÁC BẢNG.....	vi
DANH MỤC CÁC HÌNH VẼ .....	vii
DANH MỤC CÁC TỪ VIẾT TẮT.....	ix
MỞ ĐẦU .....	1
CHƯƠNG 1: TỔNG QUAN VỀ HỆ TƯ VẤN .....	1
1.1. GIỚI THIỆU TỔNG QUAN VỀ HỆ TƯ VẤN .....	1
1.2. MỘT SỐ KỸ THUẬT LỌC CỘNG TÁC.....	2
1.2.1. Lọc cộng tác lân cận.....	2
1.2.2. Lọc cộng tác phân tích ma trận.....	6
1.3. NHỮNG VẤN ĐỀ CỦA HỆ TƯ VẤN LỌC CỘNG TÁC .....	9
1.4. KẾT LUẬN.....	9
CHƯƠNG 2: TƯ VẤN DỰA TRÊN MÔ HÌNH KẾT HỢP.....	10
2.1. GIỚI THIỆU VỀ MÔ HÌNH HÀNH VI.....	10
2.1.1. Hành vi đánh giá.....	11
2.1.2. Hành vi bình luận.....	11
2.1.3. Mô hình kết hợp hành vi đánh giá và hành vi bình luận .....	11
2.2. ỨNG DỤNG MÔ HÌNH KẾT HỢP VÀO HỆ TƯ VẤN.....	12
2.2.1. Phát biểu bài toán tư vấn khách sạn dựa trên mô hình kết hợp .....	13
2.2.2. Mô hình xử lý bài toán.....	13
2.3. PHÂN LOẠI QUAN ĐIỂM NGƯỜI DÙNG.....	14
2.3.1. Tiền xử lý dữ liệu.....	14
2.3.2. Trích chọn đặc trưng .....	14
2.3.3. Mô hình học máy có giám sát cho bài toán phân loại quan điểm người dùng .....	15
2.4. KẾT LUẬN.....	19
CHƯƠNG 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ.....	20
3.1. PHÁT BIỂU BÀI TOÁN .....	20

3.2. DỮ LIỆU THỬ NGHIỆM.....	21
3.2.1. Bộ dữ liệu Booking .....	21
3.2.2. Tripadvisor Dataset .....	22
3.3. MÔI TRƯỜNG THỬ NGHIỆM.....	26
3.4. TIÊU CHÍ ĐÁNH GIÁ.....	26
3.4.1. Tiêu chí đánh giá sử dụng cho phân loại văn bản.....	26
3.4.2. Tiêu chí đánh giá sử dụng cho hệ tư vấn lọc cộng tác .....	27
3.5. KẾT QUẢ THỬ NGHIỆM .....	28
3.5.1. Bài toán 1: Khảo sát các mô hình phân loại văn bản .....	28
3.5.2. Bài toán 2: Khảo sát thuật toán lọc cộng tác.....	29
3.5.3. Bài toán 3: Khảo sát các giá trị $\alpha$ và $\beta$ trong công thức (2.3) .....	30
3.6. TỔNG HỢP ĐÁNH GIÁ.....	32
3.7. KẾT LUẬN.....	32
CHƯƠNG 4: PHÁT TRIỂN ỨNG DỤNG HỆ TƯ VẤN KHÁCH SẠN .....	33
4.1. TỔNG QUAN HỆ THỐNG .....	33
4.1.1. Giới thiệu hệ thống.....	33
4.1.2. Công nghệ sử dụng.....	33
4.2. PHÂN TÍCH HỆ THỐNG.....	34
4.2.1. Xây dựng biểu đồ ca sử dụng .....	34
4.2.2. Kịch bản ca sử dụng .....	36
4.2.3. Xây dựng biểu đồ tuần tự của các ca sử dụng.....	37
4.2.4. Xây dựng biểu đồ lớp phân tích.....	38
4.2.5. Xây dựng biểu đồ mô hình dữ liệu.....	41
4.3. THIẾT KẾ HỆ THỐNG.....	41
4.4. GIAO DIỆN MỘT SỐ CHỨC NĂNG HỆ THỐNG .....	43
4.5. KẾT LUẬN.....	43
KẾT LUẬN.....	44
DANH MỤC TÀI LIỆU THAM KHẢO .....	45
PHỤ LỤC.....	47

**DANH MỤC CÁC BẢNG**

Bảng 1.1: Ví dụ ma trận tiện ích ích	3
Bảng 2.1: Biểu diễn N-grams cho một câu	15
Bảng 3.1: Thống kê số số người dùng theo số lượng đánh giá được đưa ra	24
Bảng 3.2: Bảng phân chia các bộ dữ liệu theo các tiêu chí khác nhau	26
Bảng 3.3: Môi trường thử nghiệm	26
Bảng 3.4: Thư viện hỗ trợ chính	26
Bảng 3.5: Bảng ma trận hỗn độn	27

## DANH MỤC CÁC HÌNH VẼ

Hình 1.1: Mô hình thuật toán lọc cộng tác	2
Hình 1.2: Phân tích ma trận với K đặc trưng ẩn	7
Hình 2.1: Trang cá nhân của một người dùng trên mạng xã hội Tripadvisor	10
Hình 2.2: Mô hình xử lý bài toán tư vấn khách sạn với mô hình dữ liệu kết hợp	13
Hình 2.3: Các hàm kích hoạt	18
Hình 3.1: Sự phân bố của các phần bình luận	22
Hình 3.2: Một bản ghi trong phần thông tin khách sạn	23
Hình 3.3: Một bản ghi trong phần thông tin đánh giá từ người dùng	24
Hình 3.4: Thống kê phân bố số lượng đánh giá cho khách sạn theo khoảng	25
Hình 3.5: Phân bố điểm đánh giá	25
Hình 3.6: Biểu đồ so sánh Accuracy giữa 3 mô hình phân loại	28
Hình 3.7: Biểu đồ so sánh F1-Score giữa 3 mô hình phân loại	29
Hình 3.8: Biểu đồ so sánh MF-CF và II-CF theo tiêu chí RMSE	29
Hình 3.9: Biểu đồ so sánh MF-CF và II-CF theo tiêu chí MAE	30
Hình 3.10: Biểu đồ so sánh RMSE khi thay đổi tỷ lệ alpha-beta khi thực nghiệm MF với từng bộ dữ liệu	31
Hình 3.11: Biểu đồ so sánh MAE khi thay đổi tỷ lệ alpha-beta khi thực nghiệm MF với từng bộ dữ liệu	31
Hình 4.1: Biểu đồ tuần tự ca sử dụng đăng bài đánh giá khách sạn	37
Hình 4.2: Biểu đồ tuần tự ca sử dụng khám phá khách sạn	38
Hình 4.3: Biểu đồ lớp phân tích	40
Hình 4.4: Biểu đồ mô hình dữ liệu	41
Hình 4.5: Biểu đồ gói của hệ thống	41
Hình 4.6: Biểu đồ lớp thiết kế	42
Hình 4.7: Giao diện đăng bài đánh giá khách sạn	43
Hình 4.8: Giao diện tư vấn khách sạn	43
Hình PL. 1: Giao diện đăng nhập	47
Hình PL. 2: Giao diện trang chủ	47



Hình PL. 3: Giao diện trang cá nhân	48
Hình PL. 4: Giao diện chỉnh sửa bài viết 1	48
Hình PL. 5: Giao diện chỉnh sửa bài viết 2	49
Hình PL. 6: Giao diện kết quả tìm kiếm	49
Hình PL. 7: Giao diện chi tiết khách sạn	50
Hình PL. 8: Giao diện thích và bình luận	50

**DANH MỤC CÁC TỪ VIẾT TẮT**

<b>Từ viết tắt</b>	<b>Tiếng Anh</b>	<b>Tiếng Việt/Giải thích</b>
CF	Collaborative Filtering	Lọc cộng tác
CSS	Cascading Style Sheets	
EDI	Electronic Data Interchange	Trao đổi dữ liệu điện tử
HTML	HyperText Markup Language	
II-CF	Item-item Collaborative Filtering	Lọc cộng tác sản phẩm
IDF	Inverse Document Frequency	Nghịch đảo tần suất tài liệu
JS	Java Script	Ngôn ngữ lập trình Java Script
KNN	K-Nearest Neighbors	K láng giềng gần nhất
MF	Matrix Factorization	Lọc cộng tác phân tích ma trận
MAE	Mean Absolute Error	Sai số toàn phương trung bình
RMSE	Root Mean Square Error	Sai số toàn phương
SVM	Support Vector Machine	Máy véc-tơ hỗ trợ
TF	Term Frequency	Tần suất thuật ngữ
TMĐT		Thương mại điện tử

## MỞ ĐẦU

Cuộc sống của con người ngày càng phát triển, các nhu cầu cá nhân như: giao lưu, kết bạn, tiêu dùng, du lịch, ... ngày tăng. Nhu cầu tiêu dùng ngày càng tăng, cùng với sự phát triển của công nghệ thông tin, các hệ thống thương mại điện tử ra đời và ngày càng lớn mạnh, tiêu biểu như: Facebook, Youtube, Tripadvisor, ... Những trang thương mại điện tử này hỗ trợ doanh nghiệp quảng bá sản phẩm tới tay người tiêu dùng nhanh hơn so với bán hàng truyền thống. Tuy nhiên, khi người dùng được tiếp cận sản phẩm, dịch vụ một cách nhanh chóng thì họ cũng phải đối mặt với vấn đề có quá nhiều sản phẩm và dịch vụ và đâu thực sự là sản phẩm họ cần. Đây là tình trạng quá tải thông tin, khi người dùng có quá nhiều lựa chọn. Tuy nhiên, đôi khi họ cũng phải đối mặt với tình huống nghịch lý rằng có rất nhiều thông tin, nhưng thường rất khó để có thông tin phù hợp [1]. Với hiện trạng nêu trên, nhu cầu cấp thiết đặt ra cần có các hệ thống tự động hóa, hỗ trợ người dùng lọc thông tin cũng như cá nhân hóa đối với từng người dùng.

Hệ tư vấn ra đời nhằm giải quyết vấn đề quá tải thông tin từ người dùng, giúp họ khám phá những sản phẩm khác nhau nằm trong sở thích của mình. Có rất nhiều trang thương mại điện tử lớn sử dụng hệ tư vấn nhằm cải thiện doanh thu và tăng sự thân thiện với người dùng, một trong số đó là Youtube. Youtube, ra đời vào tháng 2, 2005 với sự phát triển nhanh chóng đã trở thành nền tảng chia sẻ video trực tuyến lớn nhất hiện nay với hơn 1 tỷ lượt xem mỗi ngày từ hàng triệu người dùng và mỗi phút có hơn 24 giờ thời lượng video được tải lên nền tảng này. Hệ tư vấn là một phần trong sự thành công của Youtube khi đóng góp 60% lượt bấm xem video từ trang chủ và các video được gợi ý từ hệ thống có tỷ lệ bấm xem gấp 2 lần những video được nhiều người xem nhất và được đánh giá cao nhất [2].

Một trong các thuật toán tư vấn điển hình và phổ biến là lọc cộng tác và hoạt động rất hiệu quả. Các hệ tư vấn truyền thống thường sử dụng dữ liệu điểm đánh giá để làm cơ sở tư vấn. Tuy nhiên, theo [3], thuật toán này vẫn còn những vấn đề tồn tại như:

- Vấn đề người dùng mới, sản phẩm mới (Cold Start)
- Vấn đề thừa thớt dữ liệu

Do thói quen lười đánh giá từ người dùng, gây ra những vấn đề trên ảnh hưởng tới độ chính xác của hệ tư vấn lọc cộng tác.

Với sự bùng nổ của các trang thương mại điện tử, các hành vi bày tỏ quan điểm ngày càng đa dạng và phong phú. Do đó, các phương pháp phân loại văn bản ngày càng được cải thiện và trở nên chính xác hơn. Những dữ liệu văn bản này cũng mang ý nghĩa bày tỏ quan điểm đối với sản phẩm.

Để hệ tư vấn có những đề xuất chính xác hơn cũng như tận dụng dữ liệu văn bản cùng các kỹ thuật phân loại được phát triển, đề án lựa chọn đề tài **“Mô hình kết hợp hành vi đánh giá và bình luận cho tư vấn khách sạn”** với mục tiêu nghiên cứu lý

thuyết về hệ tư vấn, các kỹ thuật tư vấn, tiền xử lý văn bản và phân loại văn bản về lĩnh vực cụ thể là gợi ý các khách sạn trên các bộ dữ liệu thu thập được.

Đồ án được chia thành 4 chương với nội dung như sau:

### **Chương 1: Tổng quan về hệ tư vấn**

Nội dung trong Chương 1 giới thiệu tổng quan về hệ tư vấn và các kỹ thuật lọc cộng tác. Ngoài ra, Chương 1 còn trình bày ngắn gọn các vấn đề còn tồn tại của hệ tư vấn lọc cộng tác.

### **Chương 2: Tư vấn dựa trên mô hình kết hợp**

Trong chương này, đồ án trình bày về mô hình kết hợp giữa hành vi đánh giá và hành vi bình luận và cách ứng dụng mô hình kết hợp vào hệ tư vấn lọc cộng tác. Ngoài ra, nội dung Chương 2 còn trình bày về các kỹ thuật tiền xử lý dữ liệu văn bản cùng với 3 kỹ thuật phân loại văn bản: Naïve Bayes, Logistic Regression, SVM.

### **Chương 3: Thử nghiệm và đánh giá**

Chương 3 tập trung trình bày về bộ dữ liệu được thử nghiệm, phương pháp thực nghiệm, bộ dữ liệu được sử dụng và kết quả thực nghiệm và đánh giá.

### **Chương 4: Phát triển ứng dụng hệ thống tư vấn khách sạn**

Trong Chương 4, đồ án tập giới thiệu tổng quan về hệ thống với công nghệ được sử dụng. Ngoài ra nội dung chương còn có các bước phân tích và thiết kế mô tả chi tiết cho hệ thống.

## CHƯƠNG 1: TỔNG QUAN VỀ HỆ TƯ VẤN

Trong Chương 1, đồ án trình bày một cách tổng quan về hệ tư vấn. Ngoài ra, vai trò, lợi ích của hệ tư vấn đối với thương mại điện tử cũng được trình bày trong chương này. Nội dung của Chương 1 được phân chia như sau:

- Giới thiệu tổng quan về hệ tư vấn
- Một số kỹ thuật lọc cộng tác
- Những vấn đề của hệ tư vấn lọc cộng tác
- Kết luận

### 1.1. GIỚI THIỆU TỔNG QUAN VỀ HỆ TƯ VẤN

Sự phát triển mạnh mẽ của lĩnh vực công nghệ thông tin đã góp phần giúp cuộc sống của con người ngày trở nên dễ dàng và tiện lợi. Tận dụng các thành tựu của khoa học công nghệ, nhiều trang thương mại điện tử ra đời và ngày càng lớn mạnh với sự tham gia của đông đảo người dùng tiêu biểu như: Facebook, Youtube, Netflix, Amazon, Twitter, v.v.. Thông qua các trang thương mại điện tử này, quá trình tiếp thị của những nhà cung cấp dịch vụ và sản phẩm trở nên đơn giản và dễ dàng thông qua các hình thức quảng cáo. Tuy nhiên, số lượng sản phẩm và dịch vụ ngày càng nhiều, người dùng cần phải tốn nhiều thời gian hơn trong quá lựa chọn. Đây là tình trạng quá tải thông tin, gây ra sự bất tiện và khó khăn trong quá trình trích lọc thông tin của người dùng. Ngoài ra, người dùng cũng phải đối mặt với nghịch lý rằng có rất nhiều sản phẩm để lựa chọn nhưng lại không chọn ra được một sản phẩm thích hợp.

Với hiện trạng nêu trên, hệ tư vấn ngày càng đóng vai trò quan trọng trong sự phát triển của thương mại điện tử. Theo Wikipedia, hệ tư vấn là các kỹ thuật được sử dụng nhằm mục đích dự đoán điểm đánh giá mà người dùng có thể dành cho một sản phẩm. Các điểm đánh giá dự đoán này là cơ sở để thực hiện tư vấn sản phẩm phù hợp cho người dùng. Hiện nay, các hệ thống lớn cung cấp sản phẩm, dịch vụ đều phát triển hệ tư vấn của riêng mình, tiêu biểu như: hệ tư vấn phim của Netflix, hệ tư vấn âm nhạc của Pandora, hệ tư vấn sách của Amazon [4]. Theo [4], khi sử dụng hệ tư vấn, nhà cung cấp sản phẩm và dịch vụ có thể nhận lại rất nhiều lợi ích trong đó có: tăng doanh thu bán hàng và sự hài lòng của khách hàng. Tuy nhiên, để có thể tư vấn chính xác, hệ tư vấn cần được cung cấp các dữ liệu liên quan tới sở thích và nhu cầu của người dùng. Sở thích và nhu cầu của người dùng thể hiện qua: lịch sử tìm kiếm, lịch sử mua hàng, đánh giá sản phẩm, v.v.. Những dữ liệu này đóng vai trò quyết định tới kết quả tư vấn của hệ thống.

Theo [5], các kỹ thuật sử dụng trong hệ tư vấn được chia thành 3 nhóm chính:

- **Lọc dựa trên nội dung:** Trong cách tiếp cận này, hệ thống sẽ thu thập các dữ liệu rõ ràng (điểm đánh giá sản phẩm) hoặc dữ liệu ngầm (bấm vào một đường dẫn) và tạo ra hồ sơ người dùng. Hệ thống sẽ thực hiện tư vấn những sản phẩm dựa trên những sản phẩm và hành vi liên quan tới hồ sơ người dùng. Do sở thích

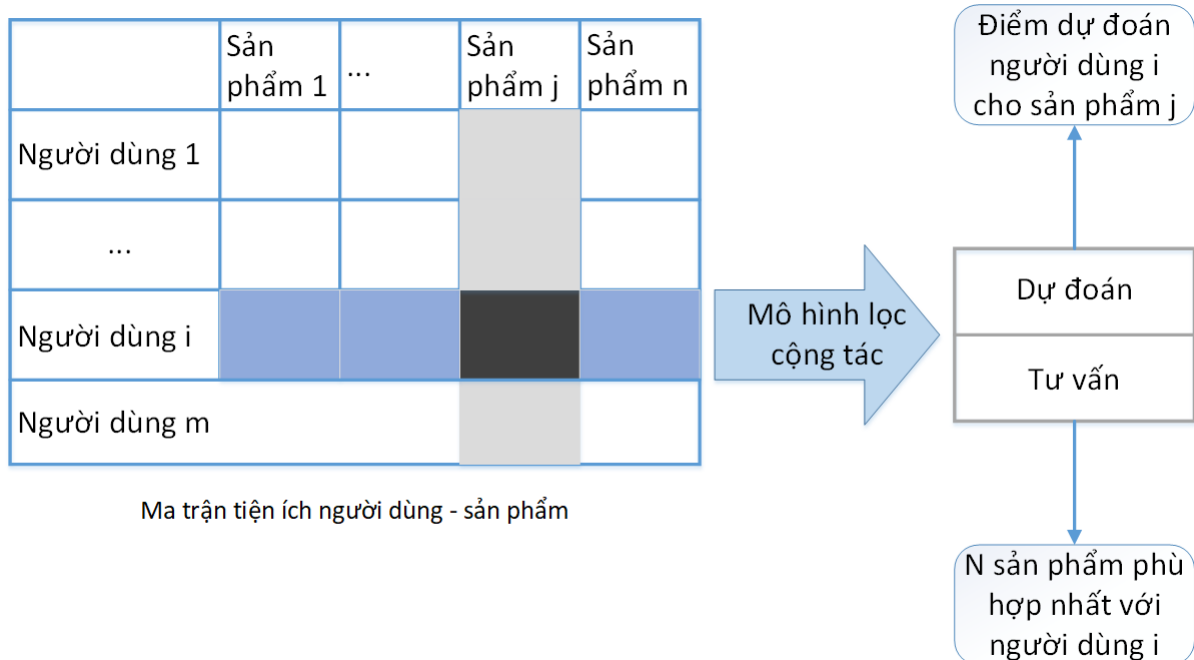
của người dùng thường được chia thành vài nhóm cơ bản, việc chỉ sử dụng hồ sơ của 1 người dùng khiến hệ thống không tận dụng được thông tin từ những người dùng khác, từ đó hạn chế sự linh hoạt của hệ tư vấn.

- **Lọc cộng tác:** Không giống với lọc dựa trên nội dung, lọc cộng tác tìm kiếm những người dùng có sở thích tương tự nhau. Từ giả định những người dùng A có sở thích giống với người dùng B, hệ thống sẽ tiến hành tư vấn cho người dùng B những sản phẩm phù hợp người dùng A. Lọc cộng tác có 2 hướng tiếp cận: dựa trên bộ nhớ và dựa trên mô hình. Hướng tiếp cận dựa trên bộ nhớ tính toán độ tương tự giữa các người dùng từ đó thực hiện tư vấn. Nhược điểm của hướng tiếp cận này là sự tổn kém tài nguyên khi số lượng người dùng và sản phẩm tăng lên. Hướng tiếp cận dựa trên mô hình sử dụng các mô hình đã được huấn luyện thông qua các thuật toán học máy hoặc khai phá dữ liệu để thực hiện tư vấn.
- **Hệ tư vấn lai:** Lọc dựa trên nội dung và lọc cộng tác đều có ưu điểm và nhược điểm riêng. Để giải quyết vấn đề này, hệ tư vấn lai được sinh ra, là sự kết hợp của 2 kỹ thuật trên.

Trong phần tiếp theo, đồ án sẽ tập trung vào việc trình bày một số kỹ thuật lọc cộng tác tiêu biểu.

## 1.2. MỘT SỐ KỸ THUẬT LỌC CỘNG TÁC

### 1.2.1. Lọc cộng tác lân cận



Hình 1.1: Mô hình thuật toán lọc cộng tác

Ý tưởng của kỹ thuật Lọc cộng tác là từ những hành vi thể hiện mối tương quan với sản phẩm, hệ thống sẽ tính toán mức độ tương đồng giữa người dùng với người dùng hoặc sản phẩm với sản phẩm, là cơ sở thực hiện tư vấn. Những người dùng có mức độ tương đồng giống nhau sẽ có xu hướng mua những sản phẩm giống nhau. Với mỗi cách

tính độ tương đồng sẽ cho một thuật toán lọc cộng tác khác. Để tính toán được mức độ tương đồng, hệ thống cần xây dựng hồ sơ cho người dùng – sản phẩm. Thông thường, hồ sơ người dùng – sản phẩm thường được xây dựng từ điểm đánh giá người dùng chấm cho sản phẩm, được gọi là ma trận tiện ích. Ma trận tiện ích sẽ có dạng như trong Hình 1.1, với các hàng/cột là danh sách người dùng, cột/hàng là danh sách sản phẩm, các giá trị trong mỗi ô tương ứng với điểm đánh giá người dùng dành cho sản phẩm. Trong thực tế, người dùng thường ít đánh giá sản phẩm nên ma trận tiện ích trở nên thưa thớt, nghĩa là có nhiều giá trị chưa được điền. Hình 1.1 là mô hình xử lý, mô tả cho thuật toán lọc cộng, tác được chia thành 3 bước thực hiện:

1. Chuẩn hóa dữ liệu
2. Tính toán độ tương đồng
3. Dự đoán mức độ quan tâm của người dùng lên sản phẩm

#### a. Lọc cộng tác người dùng

##### Chuẩn hóa dữ liệu

Trong thực tế, người dùng “lười” đánh giá sản phẩm khiến ma trận tiện ích trở nên thưa thớt. Do đó cần chuẩn hóa dữ liệu để loại bỏ những giá trị chưa biết trong ma trận. Xét ví dụ trong Bảng 1.1 là ma trận tiện ích được xây dựng từ tập người dùng  $W = \{w_1, w_2, \dots, w_5\}$  và tập sản phẩm  $X = \{x_1, x_2, \dots, x_5\}$ . Mỗi sản phẩm được người dùng đánh giá trên thang điểm từ 0 đến 5. Các giá trị “?” nghĩa là người dùng chưa đánh giá những sản phẩm tương ứng.

Bảng 1.1: Ví dụ ma trận tiện ích

	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$
$w_1$	5	5	2	0	?
$w_2$	2	4	0	?	?
$w_3$	0	1	3	4	5
$w_4$	5	?	?	?	1
$w_5$	?	?	3	2	4

Các dễ nhất để điền các giá trị còn thiếu vào trong ma trận này là chọn điểm cao nhất hoặc điểm thấp nhất (5 hoặc 0). Tuy nhiên, khi chọn giá trị này sẽ gây mất cân bằng và giảm độ chính xác của hệ thống. Một giá trị an toàn có thể điền là điểm trung bình của thang đo (2,5). Tuy nhiên, giá trị này sẽ không đúng với những người dùng khó tính hoặc dễ tính. Vì người dùng khó tính sẽ chỉ cho 4 với những sản phẩm họ thích, ngược lại người dùng dễ tính sẽ cho 1, 2 với những sản phẩm họ không thích. Do đó cần có một cách chuẩn hóa khác để khắc phục vấn đề này. Các bước chuẩn hóa được thực hiện như sau:

1. Tính trung bình các điểm đánh giá mà mỗi người dùng đã đưa ra. Ví dụ, người dùng  $w_1$  đã chấm 4 sản phẩm với số điểm lần lượt là: 5, 5, 2, 0. Như vậy, điểm trung bình người dùng  $w_1$  đưa ra là:  $\frac{5+5+2+0}{4} = 3$ .

	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	Điểm TB
$w_1$	5	5	2	0	?	3
$w_2$	2	4	0	?	?	2
$w_3$	0	1	3	4	5	2,6
$w_4$	5	?	?	?	1	3
$w_5$	?	?	3	2	4	3

2. Thực hiện trừ điểm đánh giá của người dùng với điểm đánh giá trung bình của họ.

	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	Điểm TB
$w_1$	$5 - 3 = 2$	$5 - 3 = 2$	-1	-3	?	3
$w_2$	0	2	-2	?	?	2
$w_3$	-2,6	-1,6	0,4	1,4	2,4	2,6
$w_4$	2	?	?	?	-2	3
$w_5$	?	?	0	-1	1	3

3. Các ô chưa biết giá trị thì điền 0.

	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$
$w_1$	2	2	-1	-3	0
$w_2$	0	2	-2	0	0
$w_3$	-2,6	-1,6	0,4	1,4	2,4
$w_4$	2	0	0	0	-2
$w_5$	0	0	0	-1	1

Cách chuẩn hóa này có những ưu điểm sau:

- Việc trừ đi điểm đánh giá trung bình của người dùng khiến ma trận có giá trị âm, dương. Những giá trị dương ứng với những sản phẩm được người dùng quan tâm hơn. Những ô có giá trị 0 biểu diễn cho người dùng chưa đánh giá sản phẩm này. Đây là những giá trị cần dự đoán.
- Số chiều của ma trận tiện ích là rất lớn khi người dùng và sản phẩm tăng lên. Vì vậy, để tiết kiệm bộ nhớ, ma trận tiện ích sẽ được lưu dưới dạng ma trận thưa do những dấu “?” đã được thay bằng giá trị 0.

### Tính toán độ tương đồng và dự đoán mức độ quan tâm của người dùng lên sản phẩm

Với mỗi cách tính độ tương đồng sẽ cho ra một thuật toán lọc cộng tác khác nhau. Nếu tính độ tương đồng giữa các cặp người dùng ta có thuật toán lọc cộng tác người dùng. Nếu tính độ tương đồng giữa các cặp sản phẩm, ta có thuật toán lọc cộng tác sản phẩm.

Để tính độ tương đồng giữa người dùng  $w_i$  và  $w_j$ , ta sử dụng công thức cô-sin:



$$\text{cosin\_similarity}(w_i, w_j) = \cos(w_i, w_j) = \frac{w_i^T w_j}{\|w_i\|_2 \|w_j\|_2} \quad (1.1)$$

Trong đó,  $w_i$  và  $w_j$  là các véc-tơ tương ứng với hàng/cột  $w_i$  và  $w_j$  trong ma trận tiện ích. Sau khi tính toán được độ tương đồng giữa các cặp người dùng, thuật toán sẽ dự đoán mức độ quan tâm của người dùng  $u$  lên sản phẩm  $i$  dựa trên thông tin từ  $K$  người dùng giống  $u$  nhất, được định nghĩa theo công thức:

$$\hat{y}_{i,u} = \frac{\sum_{u_j \in N(u,i)} \bar{y}_{i,u_j} \text{sim}(u, u_j)}{\sum_{u_j \in N(u,i)} |\text{sim}(u, u_j)|} \quad (1.2)$$

Trong đó,  $N(u, i)$  là tập hợp  $K$  người dùng gần giống  $u$  nhất và đã đánh giá sản phẩm  $i$ .

Xét ví dụ đã được trình bày trong Chuẩn hóa dữ liệu, dự đoán độ quan tâm của  $w_1$  lên  $x_5$  sử dụng lọc cộng tác người dùng.

- Người dùng đã đánh giá  $x_5$ :  $\{w_3, w_4\}$
- Độ tương tự tương ứng giữa  $w_1$  và  $w_3$ :  

$$\frac{2 * (-2,6) + 2 * (-1,6) + (-1) * 0,4 + (-3) * 1,4 + 0 * 2,4}{\sqrt{2^2 + 2^2 + (-1)^2 + (-3)^2 + 0^2} * \sqrt{(-2,6)^2 + (-1,6)^2 + 0,4^2 + 1,4^2 + 2,4^2}} = -0,7$$
- Độ tương tự giữa  $w_1$  và  $w_4$ :  $\frac{1}{3}$
- Xét  $K=2$ , 2 người dùng giống  $w_1$  nhất:  $N(u, i) = \{w_3, w_4\}$  với điểm đánh giá chuẩn hóa là 2,4, -2.
- $\hat{y}_{w_1, x_5} = \frac{-2 * \frac{1}{3} + 2,4 * -0,7}{\frac{1}{3} + |-0,7|} = 0,981$
- Sau đó, để đưa điểm đánh giá về thang đo ban đầu, ta cộng điểm đánh giá dự đoán với điểm đánh giá trung bình của người dùng ta có:  $0,981 + 3 = 3,981$ .

Lọc cộng tác người dùng thường hoạt động không hiệu quả trên các hệ thống lớn do số lượng người dùng khổng lồ. Khi đó, việc tính toán độ tương đồng giữa các cặp người dùng trở nên tốn kém tài nguyên là thời gian.

### b. Lọc cộng tác sản phẩm

Lọc cộng tác sản phẩm là hướng tiếp cận có thể khắc phục nhược điểm của lọc cộng tác người dùng do số lượng sản phẩm trên hệ thống thường không biến động mạnh. Thay vì tính toán độ tương đồng giữa các cặp người dùng, lọc cộng tác sản phẩm tính toán độ tương đồng giữa các sản phẩm.

### Chuẩn hóa dữ liệu

#### 1. Tính trung bình điểm đánh giá sản phẩm nhận được

	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$
$w_1$	5	5	2	0	?

$w_2$	2	4	0	?	?
$w_3$	0	1	3	4	5
$w_4$	5	?	?	?	1
$w_5$	?	?	3	2	4
<b>Điểm TB</b>	3	3,333	2	2	3,333

2. Thực hiện trừ điểm đánh giá của sản phẩm với điểm đánh giá trung bình.

	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$
$w_1$	2	1,667	0	-2	?
$w_2$	-1	0,667	-2	?	?
$w_3$	-3	-2,333	1	2	1,667
$w_4$	2	?	?	?	-2,333
$w_5$	?	?	1	0	0,667
<b>Điểm TB</b>	3	3,333	2	2	3,333

3. Các ô “?” điền giá trị 0

	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$
$w_1$	2	1,667	0	-2	0
$w_2$	-1	0,667	-2	0	0
$w_3$	-3	-2,333	1	2	1,667
$w_4$	2	0	0	0	-2,333
$w_5$	0	0	1	0	0,667

**Tính toán độ tương đồng và dự đoán mức độ quan tâm của người dùng lên sản phẩm**

Dự đoán độ quan tâm của  $w_2$  lên  $x_5$  sử dụng lọc cộng tác sản phẩm.

- Sản phẩm được  $w_2$  đánh giá:  $\{x_1, x_2, x_3\}$
- Độ tương tự tương ứng giữa  $x_5$  và  $\{x_1, x_2, x_3\}$  lần lượt là:  $\{-0,774, -0,449, 0,324\}$
- Xét  $K=2$ , 2 sản phẩm giống  $x_5$  nhất:  $N(u, i) = \{x_2, x_3\}$  với điểm đánh giá chuẩn hóa là 0,667, -2.
- $\hat{y}_{w_2, x_5} = \frac{0,667 \cdot -0,449 + (-2) \cdot 0,324}{0,324 + |-0,449|} = -1,226$
- Đưa điểm đánh giá về thang đo ban đầu, ta cộng điểm đánh giá dự đoán với điểm đánh giá trung bình của sản phẩm ta có:  $-1,226 + 3,333 = 2,107$ .

### 1.2.2. Lọc cộng tác phân tích ma trận

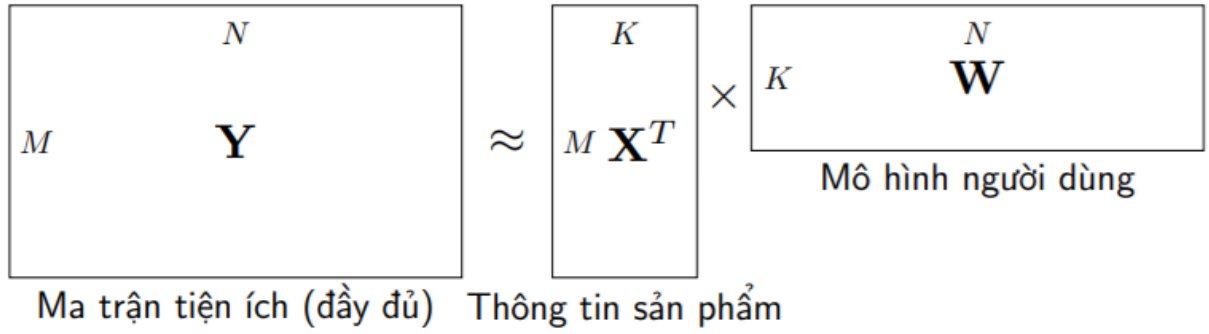
#### Giới thiệu

Ý tưởng chính của phương pháp này là tồn tại các đặc trưng ẩn mô tả sự liên quan giữa các sản phẩm và người dùng. Ví dụ với các bộ phim, các đặc trưng ẩn có thể rõ

ràng như: hài, chính kịch, hành động, hoặc chúng là sự kết hợp của các đặc trưng ẩn rõ ràng, hoặc chúng là những đặc trưng chưa được đặt tên. Tương tự, mỗi người dùng cũng sẽ có xu hướng thích những đặc trưng ẩn nào đó của phim. Thay vì xây dựng ma trận của  $M$  sản phẩm  $X$  một cách độc lập, các đặc trưng ẩn này được huấn luyện đồng thời với dữ liệu của ma trận  $N$  người dùng  $Y$ .

Với ý tưởng trên, thay vì xây dựng ma trận  $Y$  nghĩa là dự đoán các giá trị còn khuyết trong  $Y$  thì thuật toán sẽ cố gắng sắp xếp ma trận người dùng  $W$  và ma trận sản phẩm  $X$ , sao cho tích của 2 ma trận này là  $\hat{Y}$  xấp xỉ với  $Y$ .

$$Y \approx \hat{Y} = X^T W \quad (1.3)$$



Hình 1.2: Phân tích ma trận với  $K$  đặc trưng ẩn [6]

$K$  là số đặc trưng ẩn được giả định của mỗi sản phẩm. Thông thường,  $K$  được chọn là một số nhỏ hơn  $M$  và  $N$  rất nhiều. Khi đó, hạng của  $X$  và  $Y$  không cao, giúp tiết kiệm tài nguyên.

### Hàm mất mát

Hàm mất mát được xây dựng dựa trên các ô đã được điền của ma trận  $Y$ , được định nghĩa như sau:

$$L(X, W) = \frac{1}{2s} \sum_{n=1}^N \sum_{m:r_{mn}=1} (y_{mn} - x_m w_n)^2 + \frac{\lambda}{2} (||X||_F^2 + ||W||_F^2) \quad (1.4)$$

Trong đó  $r_{mn} = 1$  nếu sản phẩm thứ  $m$  đã được đánh giá với người dùng thứ  $n$ ,  $||\blacksquare||_F^2$  là căn bậc 2 của tổng bình phương tất cả các phần tử của ma trận,  $s$  là toàn bộ số đánh giá đã có. Trong công thức trên, thành phần thứ nhất chính là trung bình sai số của mô hình, thành phần thứ hai là  $l_2$  regularization, giúp tránh overfitting.

Việc tối ưu cả 2 ma trận  $X$  và  $W$  cùng lúc là tương đối phức tạp, vì vậy, phương pháp được sử dụng là tối ưu từng ma trận trong khi ma trận kia cố định đến khi hội tụ.

### Tối ưu hàm mất mát

Gradient Descent là kỹ thuật được dùng để tối ưu 2 bài toán: cố định  $X$  tối ưu  $W$  và cố định  $W$  tối ưu  $X$ .

*a. Cố định X tối ưu W*

Hàm mất mát:

$$L(W) = \frac{1}{2s} \sum_{n=1}^N \sum_{m:r_{mn}=1} (y_{mn} - x_m w_n)^2 + \frac{\lambda}{2} \|W\|_F^2 \quad (1.5)$$

Việc tối ưu công thức trên có thể được tách thành N bài toán nhỏ, mỗi bài toán ứng với việc đi tối ưu một cột của ma trận W:

$$L(w_n) = \frac{1}{2s} \sum_{m:r_{mn}=1} (y_{mn} - x_m w_n)^2 + \frac{\lambda}{2} \|w_n\|_2^2 \quad (1.6)$$

Vì biểu thức chỉ phụ thuộc vào các sản phẩm đã được đánh giá bởi người dùng đang xét, công thức có thể được đơn giản bằng cách đặt  $\hat{X}_n$  là ma trận được tạo bởi các hàng tương ứng với các sản phẩm đã được đánh giá đó, và  $\hat{y}_n$  là các đánh giá tương ứng. Khi đó công thức trở thành:

$$L(w_n) = \frac{1}{2s} \|\hat{y}_n - \hat{X}_n w_n\|^2 + \frac{\lambda}{2} \|w_n\|_2^2 \quad (1.7)$$

và đạo hàm của nó:

$$\frac{\partial L(w_n)}{\partial w_n} = -\frac{1}{s} \hat{X}_n^T (\hat{y}_n - \hat{X}_n w_n) + \lambda w_n \quad (1.8)$$

Từ đó, công thức cập nhật cho mỗi cột của W được định nghĩa như sau:

$$w_n = w_n - \eta \left( -\frac{1}{s} \hat{X}_n^T (\hat{y}_n - \hat{X}_n w_n) + \lambda w_n \right) \quad (1.9)$$

*b. Cố định W tối ưu X*

Hàm mất mát:

$$L(X) = \frac{1}{2s} \sum_{n=1}^N \sum_{m:r_{mn}=1} (y_{mn} - x_m w_n)^2 + \frac{\lambda}{2} \|X\|_F^2 \quad (1.10)$$

Việc tối ưu công thức trên có thể được tách thành M bài toán nhỏ, mỗi bài toán ứng với việc đi tối ưu một cột của ma trận X:

$$L(x_m) = \frac{1}{2s} \sum_{n:r_{mn}=1} (y_{mn} - x_m w_n)^2 + \frac{\lambda}{2} \|x_m\|_2^2 \quad (1.11)$$

Vì biểu thức chỉ phụ thuộc vào các sản phẩm đã được đánh giá bởi người dùng đang xét, công thức có thể được đơn giản bằng cách đặt  $\hat{W}_m$  là ma trận được tạo bởi các hàng tương ứng với các sản phẩm đã được đánh giá đó, và  $\hat{y}^m$  là các đánh giá tương ứng. Khi đó công thức trở thành:

$$L(x_m) = \frac{1}{2s} \|\hat{y}^m - x_m \hat{W}_m\|_2^2 + \frac{\lambda}{2} \|x_m\|_2^2 \quad (1.12)$$

và đạo hàm của nó:

$$\frac{\partial L(x_m)}{\partial x_m} = -\frac{1}{s}(\hat{y}^m - x_m \hat{W}_m) \hat{W}_m^T + \lambda x_m \quad (1.13)$$

Từ đó, công thức cập nhật cho mỗi cột của W được định nghĩa như sau:

$$x_m = x_m - \eta \left( -\frac{1}{s}(\hat{y}^m - x_m \hat{W}_m) \hat{W}_m^T + \lambda x_m \right) \quad (1.14)$$

### 1.3. NHỮNG VẤN ĐỀ CỦA HỆ TƯ VẤN LỘC CỘNG TÁC

Theo [3], hệ tư vấn lộc cộng tác là kỹ thuật được sử dụng phổ biến hiện nay nhưng vẫn còn phải đối mặt với những vấn đề điển hình như: khởi đầu lạnh, thừa thớt dữ liệu và khả năng mở rộng.

Đầu tiên, vấn đề thừa thớt dữ liệu, một trong những vấn đề chính của hệ tư vấn và ảnh hưởng rất nhiều đến chất lượng của hệ thống. Thông thường, dữ liệu để thực hiện tư vấn của hệ thống được biểu diễn dưới dạng ma trận người dùng-sản phẩm, giá trị của các ô trong ma trận là điểm đánh giá người dùng dành cho sản phẩm đó. Tuy nhiên, do thói quen lười đánh giá của người dùng khiến mật độ các giá trị được điền của ma trận trở nên thừa thớt. Sự thừa thớt này càng ngày càng tăng lên khi hệ thống phát triển, số lượng người dùng và sản phẩm tăng lên. Đây vẫn là một vấn đề cần phải được nghiên cứu thêm.

Tiếp theo, vấn đề khởi đầu lạnh xảy ra khi gặp 1 trong 3 tình huống: người dùng mới, sản phẩm mới và hệ thống mới. Trong những tình huống này, người dùng, sản phẩm hay hệ thống chưa có dữ liệu để thực hiện khai thác, dự đoán thói quen, nhu cầu của người dùng. Vì vậy hệ thống rất khó để thực hiện tư vấn.

Cuối cùng, khả năng mở rộng là thuộc tính của hệ thống cho thấy khả năng xử lý lượng thông tin ngày càng tăng một cách hiệu quả. Với sự bùng nổ dữ liệu, đây là một thách thức lớn đối với các hệ thống khi nhu cầu xử lý thông tin liên tục tăng. Trong lộc cộng tác, các phép tính phát triển theo cấp số nhân và tốn kém tài nguyên, đôi khi dẫn đến kết quả không chính xác.

### 1.4. KẾT LUẬN

Nội dung Chương 1 đã trình bày một cách tổng quan của hệ tư vấn, lợi ích, tầm quan trọng của kỹ thuật này trong thương mại điện tử. Ngoài ra, Chương 1 còn trình bày về kỹ thuật lộc cộng tác được sử dụng phổ biến trong các hệ tư vấn hiện nay. Mang trong mình ưu điểm khi có thể tận dụng thông tin của toàn bộ người dùng trong hệ thống để thực hiện tư vấn một cách linh hoạt nhưng đây cũng chính là điểm dẫn đến những khó khăn khi sử dụng kỹ thuật này. Trong chương tiếp theo, đồ án sẽ trình bày về mô hình kết hợp giữa dữ liệu hành vi đánh giá và dữ liệu hành vi bình luận, là cơ sở để hệ tư vấn thực hiện gợi ý.

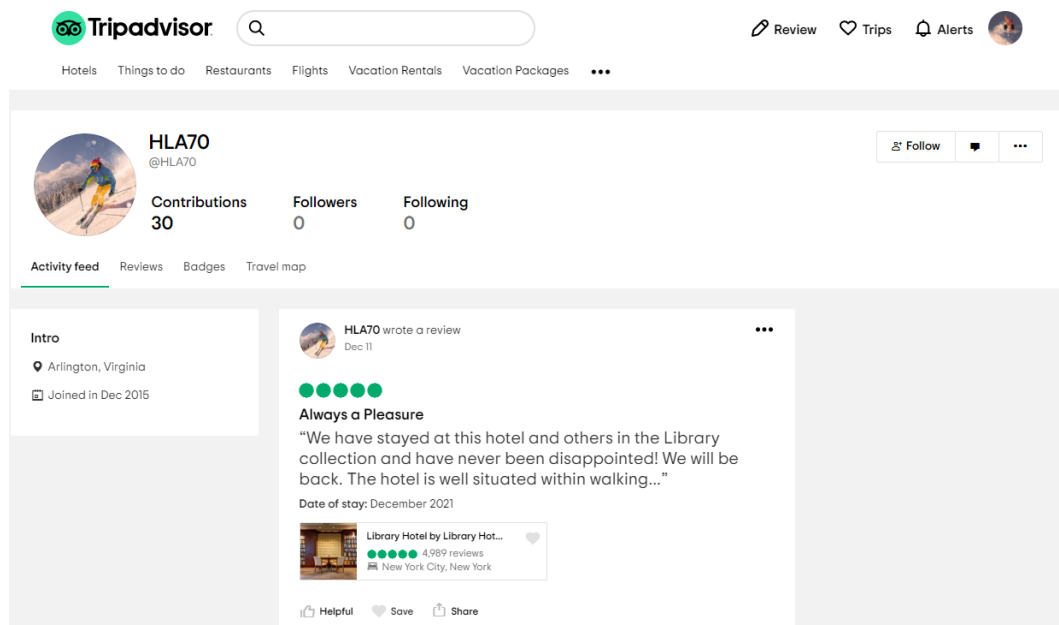
## CHƯƠNG 2: TƯ VẤN DỰA TRÊN MÔ HÌNH KẾT HỢP

Trong Chương 2, đồ án tập trung trình bày về mô hình kết hợp hành vi đánh giá và bình luận, cách ứng dụng mô hình này vào hệ tư vấn. Ngoài ra các kỹ thuật tiền xử lý dữ liệu văn bản, phân loại văn bản được đồ án sử dụng cũng được trình bày trong chương này. Nội dung Chương 2 gồm:

- Giới thiệu về mô hình hành vi
- Ứng dụng mô hình kết hợp vào hệ tư vấn
- Phân loại quan điểm người dùng

### 2.1. GIỚI THIỆU VỀ MÔ HÌNH HÀNH VI

Trong mạng xã hội, mỗi người dùng có một không gian riêng và có thể kết nối với nhau thông qua danh sách bạn bè. Trong không gian này, người dùng có quyền làm những gì họ muốn trong phạm vi hỗ trợ của nền tảng mạng xã hội, chẳng hạn như: chia sẻ một bộ phim, bình luận về một bài viết, kết bạn và theo dõi. Những hành động trên được gọi chung là hành vi của người dùng trên mạng xã hội. Các hành vi của người dùng trên mạng xã hội phản ánh một phần sở thích, tính cách và quan điểm của họ đối với những sự kiện xảy ra trên mạng xã hội. Điều này có ảnh hưởng không nhỏ tới những người trong danh sách bạn bè của họ.



Hình 2.1: Trang cá nhân của một người dùng trên mạng xã hội Tripadvisor

Hình 2.1 mô tả trang cá nhân của một người dùng trên mạng xã hội Tripadvisor. Tripadvisor là một trang chuyên cung cấp thông tin về những địa điểm du lịch: nhà hàng, khách sạn, danh lam thắng cảnh. Những người dùng trên nền tảng này để lại đánh giá cho những địa điểm mà họ đã trải nghiệm. Những đánh giá này ảnh hưởng tới quyết định trải nghiệm du lịch của những người dùng khác. Càng có nhiều người theo dõi thì mức độ ảnh hưởng của người dùng càng lớn, thể hiện thông qua: số lượng người theo

dôi (Followers), tương tác của bài đánh giá (Helpful, Save, Share). Đăng bài đánh giá, theo dõi, tương tác là những hành vi chính trên mạng xã hội này.

Với một bài đánh giá, phần đánh giá điểm và bình luận là 2 phần thể hiện rõ nhất quan điểm của người dùng. Vì vậy, trong phần tiếp theo, đồ án sẽ tập trung trình bày về hành vi đánh giá và hành vi bình luận của người dùng trên mạng xã hội.

### 2.1.1. Hành vi đánh giá

Các hệ tư vấn thường được xây dựng từ:

- Tập người dùng  $W = \{w_1, w_2, \dots, w_n\}$
- Tập sản phẩm  $X = \{x_1, x_2, \dots, x_m\}$

Hành vi đánh giá là hành động người dùng chấm điểm cho sản phẩm. Thông tin này được lưu trữ và thường được sử dụng làm cơ sở cho hệ thống thực hiện tư vấn. Điểm đánh giá từ người dùng  $w_i$  cho sản phẩm  $x_j$  được định nghĩa như sau:

$$rating_{ij} = y, y \in \{1, 2, \dots, t\} \quad (2.1)$$

Trong đó,  $t$  thường được chọn là 5 hoặc 10.

### 2.1.2. Hành vi bình luận

Hành vi bình luận là hành động của người dùng khi diễn đạt suy nghĩ, quan điểm của mình bằng văn bản. Người dùng thực hiện hành vi bình luận đối với sản phẩm thay vì hành vi chấm điểm sẽ mô tả rõ hơn trải nghiệm, suy nghĩ của họ đối với sản phẩm. Mỗi bình luận  $comment_{ij}$  người dùng  $w_i$  bày tỏ quan điểm đối với sản phẩm  $x_j$ . Bình luận mang nhãn 0 nếu người dùng thích hoặc khen khách sạn. Ngược lại, bình luận mang nhãn 1 nếu người dùng không thích hoặc chê khách sạn.

$$comment_{ij} = \begin{cases} 0 & \text{nếu } w_i \text{ thích } x_j \\ 1 & \text{nếu ngược lại} \end{cases} \quad (2.2)$$

### 2.1.3. Mô hình kết hợp hành vi đánh giá và hành vi bình luận

Để sử dụng dữ liệu hành vi đánh giá và bình luận cùng lúc cho tư vấn khách sạn thì cần có một phương pháp để kết hợp hai loại dữ liệu này. Như đã trình bày trong Phần 1.2:

- $rating_{ij}$ : Điểm đánh giá từ người dùng  $w_i$  cho sản phẩm  $x_j$
- $comment_{ij}$ : Bình luận bày tỏ quan điểm từ người dùng  $w_i$  cho sản phẩm  $x_j$

Theo [7], các bình luận tiêu cực có ảnh hưởng không nhỏ tới quyết định mua hàng của người dùng. Tuy nhiên, nếu sản phẩm có nhiều phản hồi tích cực thì cũng làm tăng khả năng mua hàng của người dùng. Do đó, đồ án thực hiện kết hợp dựa trên ý tưởng: **“Nếu khách sạn có bình luận tiêu cực thì điểm đánh giá dành cho khách sạn này cần phải hạ xuống. Tuy nhiên, khách sạn có nhiều phản hồi tích cực thì điểm đánh giá cũng cần được tăng lên”**. Điều này có nghĩa là, nếu khách sạn có nhiều phản hồi tích cực thì các điểm đánh giá dành cho khách sạn này sẽ được thưởng thêm và ngược lại, nếu khách sạn có nhiều bình luận phản nản thì điểm đánh giá sẽ bị trừ đi.



Với ý tưởng trên, điểm đánh giá dự đoán  $rating_{ij}$  của người dùng  $w_i$  với khách sạn  $x_j$ , tỷ lệ số bình luận tích cực  $p\_rate_j$ , tỷ lệ số bình luận tiêu cực  $n\_rate_j$  của khách sạn  $x_j$  sẽ là 3 thành phần quyết định tới điểm đánh giá cuối cùng dành cho khách sạn. Coi điểm đánh giá cuối cùng là 100%,  $\alpha$  và  $\beta$  là 2 trọng số tương ứng của  $rating_{ij}$ ,  $p\_rate_j$  và  $n\_rate_j$  quyết định mức độ ảnh hưởng của 2 thành phần này lên điểm đánh giá cuối cùng. Đồ án biểu diễn ý tưởng thông qua công thức:

$$c\_rating(w_i, x_j) = \alpha \times rating_{ij} + \beta \times (p\_rate_j - n\_rate_j) \quad (2.3)$$

Trong đó:

- $c\_rating(w_i, x_j)$  là điểm đánh giá kết hợp, được sử dụng để làm dữ liệu thực hiện huấn luyện và đánh giá,  $c\_rating(w_i, x_j) \in [0; 5]$ .
- $rating_{ij}$  là điểm đánh giá được dự đoán thông qua thuật toán lọc cộng tác,  $rating_{ij} \in [0; 5]$ .
- $p\_rate_j$  là tỉ lệ số  $comment_{ij} = 0$  trong tổng số các  $comment_{ij}$  của khách sạn  $x_j$ ,  $positive\_rate_j \in [0; 1]$ .
- $n\_rate_j$  là tỉ lệ số  $comment_{ij} = 1$  trong tổng số các  $comment_{ij}$  của khách sạn  $x_j$ ,  $n\_rate_j \in [0; 1]$ .
- Do  $rating_{ij}$  nằm trong khoảng giá trị khác với  $p\_rate_j$  và  $n\_rate_j$ , vì vậy, trước khi thực hiện kết hợp, đồ án thực hiện chuyển  $rating_{ij}$  về cùng khoảng giá trị  $[0; 1]$  với  $p\_rate_j$  và  $n\_rate_j$  bằng cách  $rating_{ij} = \frac{rating_{ij}}{5}$ .
- $\alpha + \beta = 1$ , dùng để biểu diễn cho mức độ quan trọng của từng phần.
- Sau khi thực hiện tính toán, để đưa điểm đánh giá dự đoán về khoảng ban đầu, đồ án thực hiện  $c\_rating(w_i, x_j) = c\_rating(w_i, x_j) \times 5$ .

Để tìm ra cặp  $\alpha, \beta$  phù hợp, đồ án thực hiện khảo sát trên bộ dữ liệu khách sạn được trình bày trong Phần 3.3, với các tỷ lệ khác nhau. Kết quả khảo sát được trình bày trong Phần 3.4.3. Trong phần tiếp theo, đồ án sẽ trình bày phương pháp áp dụng mô hình kết hợp vào hệ tư vấn.

## 2.2. ỨNG DỤNG MÔ HÌNH KẾT HỢP VÀO HỆ TƯ VẤN

Theo khảo sát [8], ngày càng có nhiều bài báo được công bố có chủ đề liên quan tới hệ tư vấn. Điều này chứng tỏ các phương pháp tư vấn ngày càng được cải tiến nhưng vẫn còn nhiều vấn đề còn tồn tại. Như đã trình bày trong Phần 1.3, vấn đề khởi đầu lạnh và thừa thớt dữ liệu đánh giá làm giảm độ chính xác của các hệ thống tư vấn. Hơn nữa, những điểm số mà người dùng đánh giá cho sản phẩm đôi khi chưa phản ánh chính xác chất lượng của sản phẩm/dịch vụ do điểm số người dùng có thể chấm chỉ là các số nguyên hay do suy nghĩ chủ quan từ người dùng. Ví dụ, trên thang điểm từ 0 đến 5, người dùng khó tính thường cho 3, 4 với sản phẩm/dịch vụ họ thích và ngược lại, người dùng dễ tính thường cho 2, 3 với sản phẩm/dịch vụ họ không thích [6]. Như vậy, cần có một phương pháp cải thiện vấn đề này.



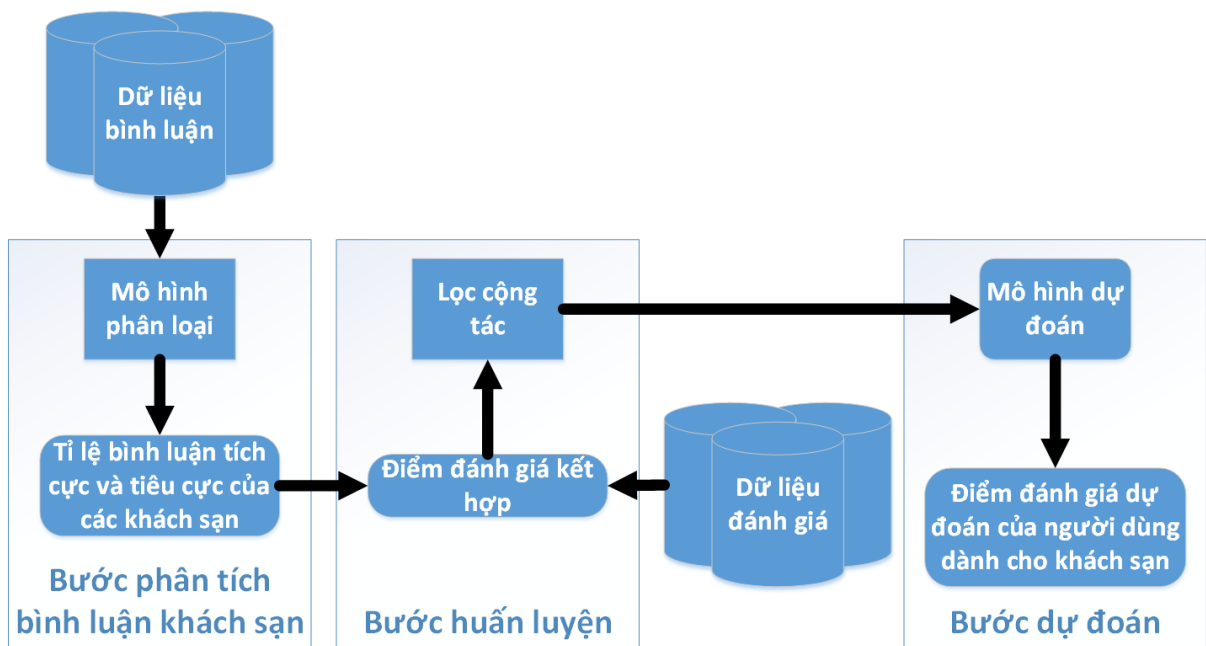
### 2.2.1. Phát biểu bài toán tư vấn khách sạn dựa trên mô hình kết hợp

Bài toán tư vấn khách sạn dựa trên hành vi đánh giá và bình luận sẽ giải quyết vấn đề làm thế nào để tư vấn khách sạn cho người dùng một cách chính xác hơn. Thay vì chỉ sử dụng một loại dữ liệu là các hành vi đánh giá, đồ án sử dụng dữ liệu hành vi đánh giá kết hợp với dữ liệu hành vi bình luận. Bài toán được phát biểu như sau:

- **Input:** Dữ liệu hành vi đánh giá và bình luận của người dùng dành cho các khách sạn
- **Output:** Tư vấn khách sạn dành cho người dùng.

### 2.2.2. Mô hình xử lý bài toán

Hình 2.2 mô tả quá trình xử lý bài toán tư vấn khách sạn với dữ liệu kết hợp. Trong đó, quá trình xử lý gồm 3 bước: phân tích bình luận khách sạn, huấn luyện và dự đoán.



Hình 2.2: Mô hình xử lý bài toán tư vấn khách sạn với mô hình dữ liệu kết hợp

#### Bước 1: Phân tích bình luận khách sạn

Mục tiêu của bước này là tính toán tỷ lệ bình luận tích cực  $p\_rate_j$  và  $n\_rate_j$  của khách sạn  $x_j$ . Để làm được điều này, đồ án sử dụng mô hình phân loại văn bản đã được huấn luyện để gán nhãn cho các bình luận của khách sạn. Sau khi các bình luận được gán nhãn, đồ án thực hiện tính  $p\_rate_j$  và  $n\_rate_j$ . Ví dụ khách sạn  $x_1$  có tất cả 10 bình luận, trong đó 7 bình luận được gán nhãn tích cực và còn lại, 3 bình luận được gán nhãn tiêu cực thì  $p\_rate_1 = 0,7$  và  $n\_rate_1 = 0,3$ .

#### Bước 2: Huấn luyện

Sau khi thực hiện phân tích bình luận khách sạn, đồ án thực hiện tính toán lại các điểm đánh giá trong bộ dữ liệu bằng cách sử dụng công thức (2.3). Ví dụ, người dùng

$w_i$  chấm điểm  $rating_{ij} = 4$  trên thang điểm từ 0 đến 5 cho khách sạn  $x_j$  có  $p\_rate_j = 0,7$  và  $n\_rate_j = 0,3$ . Điểm đánh giá kết hợp được tính toán như sau:

1. Đưa  $rating_{ij}$  trong khoảng về cùng khoảng giá trị  $[0, 1]$  với  $p\_rate_j$  và  $n\_rate_j$ .  
Khi đó,  $rating_{ij} = \frac{4}{5} = 0,8$ .
2. Sử dụng công thức (2.3):  $c\_rating(w_i, x_j) = \alpha \times 0,8 + \beta \times (0,7 - 0,3)$
3. Sau đó, thực hiện:  $c\_rating(w_i, x_j) = c\_rating(w_i, x_j) \times 5$  để đưa điểm đánh giá về thang ban đầu.

Các điểm đánh giá sau khi thực hiện tính toán lại là đầu vào của thuật toán lọc cộng tác. Kết quả của bước này là một mô hình dự đoán điểm đánh giá, là cơ sở để thực hiện tư vấn.

### Bước 3: Dự đoán

Với mô hình dự đoán đã được huấn luyện sau khi hoàn thành bước 2, hệ thống có thể dự đoán điểm đánh giá của người dùng dành cho các khách sạn mà họ chưa thực hiện đánh giá. Các kết quả dự đoán này là cơ sở để thực hiện tư vấn. Hệ thống sẽ tư vấn cho người dùng những khách sạn theo thứ tự điểm đánh giá dự đoán giảm dần.

## 2.3. PHÂN LOẠI QUAN ĐIỂM NGƯỜI DÙNG

### 2.3.1. Tiền xử lý dữ liệu

Quá trình tiền xử lý dữ liệu gồm 4 bước:

1. Chuẩn hóa văn bản: Bước này, văn bản được đưa về chữ thường, các biểu tượng cảm xúc, đường dẫn bị loại bỏ
2. Tách từ và loại bỏ dấu câu: Tách từ là đưa câu bình luận về dạng 1 danh sách các từ cũng với đó là loại bỏ các dấu câu. Các dấu câu không có ý nghĩa cho việc phân loại quan điểm.
3. Loại bỏ Stopword: Stopword là những từ xuất hiện nhiều nhưng không có ý nghĩa trong quá trình phân loại quan điểm. Ví dụ: “is”, “a”, “the”, ...
4. Chuyển về dạng chuẩn: Ví dụ: “rooms”=>”room”, “person”=>”people”, các từ được đưa về dạng nguyên bản.

### 2.3.2. Trích chọn đặc trưng

#### TF-IDF

TF-IDF là một phương pháp thống kê, nhằm phản ánh độ quan trọng của mỗi từ hoặc 1 cụm N-grams đối với văn bản trong phạm vi toàn bộ tài liệu đầu vào.

Cho một kho gồm  $p$  văn bản khác nhau  $D = \{D_1, D_2, \dots, D_N\}$ , mỗi văn bản  $D_i$  được tạo bởi các từ  $D_i = \{d_{i1}, \dots, d_{in}\}$ . Cho  $T = \{t_1, t_2, \dots, t_q\}$  là tập hợp những từ xuất hiện trong kho văn bản.

$$tf - idf(t_j, D_i) = tf(t_j, D_i) \times idf(t_j, D) \quad (2.4)$$

Trong đó,  $tf(t, d)$  của từ  $t$  trong văn bản  $d$  được định nghĩa như sau:

$$tf(t, d) = \frac{\text{số lần } t \text{ xuất hiện trong } d}{\text{số từ trong } d} \quad (2.5)$$

$idf(t_j, D)$  được định nghĩa theo công thức:

$$idf(t) = \log \left( \frac{N}{1 + |\{D_i | d \in D_i\}|} \right) \quad (2.6)$$

## N-grams

Bảng 2.1: Biểu diễn N-grams cho một câu

Bậc	Cụm từ
1 gram	That, picture, is, beautiful
2 gram	That picture, picture is, is beautiful
3 gram	That picture is, picture is beautiful

Một cụm N-grams là một dãy gồm N ký tự hoặc từ liên tiếp nhau trong một văn bản cho trước. Số phần tử trong một cụm N-grams được gọi là bậc của N-grams. Thông thường, bậc của N-grams thường nằm trong khoảng (1,3), với các tên gọi tương ứng là unigram (bậc 1), bigram (bậc 2) và trigram (bậc 3). N-grams được dùng để tính tần suất xuất hiện của 1 cụm N-grams có trong kho văn bản. Bảng 2.1 là ví dụ biểu diễn N-grams với bậc 1, 2, 3 cho câu: "That picture is beautiful."

### 2.3.3. Mô hình học máy có giám sát cho bài toán phân loại quan điểm người dùng

#### 2.3.3.1. Naïve Bayes Classifier

Bộ phân loại Bayes là một giải thuật thuộc lớp giải thuật phân lớp thống kê, nó có thể dự đoán xác suất của một phần tử dữ liệu thuộc vào một lớp là bao nhiêu. Bộ phân loại Bayes được dựa trên định lý Bayes [9].

##### a. Định lý Bayes

Theo Wikipedia, định lý Bayes cho phép tính xác suất xảy ra của một sự kiện ngẫu nhiên A khi biết sự kiện liên quan B đã xảy ra. Xác suất này được ký hiệu là  $P(A|B)$ , và đọc là "xác suất của A nếu có B". Đại lượng này được gọi là xác suất có điều kiện hay xác suất hậu nghiệm vì nó được rút ra từ giá trị được cho của B hoặc phụ thuộc vào giá trị đó. Theo định lý Bayes, xác suất xảy ra A khi biết B phụ thuộc vào 3 yếu tố:

- Xác suất xảy ra của A của riêng nó, không quan tâm đến B, ký hiệu là  $P(A)$ , đọc là xác suất của A. Đây là xác suất biên duyên hay xác suất tiên nghiệm, nó là "tiên nghiệm" nghĩa rằng nó không quan tâm tới bất cứ thông tin nào của B.
- Xác suất xảy ra B của riêng nó, không quan tâm đến A, ký hiệu là  $P(B)$  và đọc là "xác suất của B". Đại lượng này còn gọi là hằng số chuẩn hóa (Normalising Constant), vì nó luôn giống nhau, không phụ thuộc vào sự kiện A đang muốn biết.
- Xác suất xảy ra B khi biết A xảy ra. Ký hiệu là  $P(B|A)$  và đọc là "xác suất của B nếu có A". Đại lượng này gọi là khả năng (Likelihood) xảy ra B khi biết A đã xảy ra.

Khi đó, xác suất của A khi biết B được định nghĩa bởi công thức:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (2.7)$$

#### b. Bộ phân loại Naïve Bayes

Bộ phân loại Naïve Bayes hoạt động như sau:

1. Gọi D là tập dữ liệu huấn luyện, trong đó mỗi phần tử dữ liệu X được biểu diễn bằng một vector chứa n giá trị thuộc tính  $A_1, A_2, \dots, A_n, X = \{x_1, x_2, \dots, x_n\}$ .
2. Giả sử có m lớp  $C_1, C_2, \dots, C_n$ ; Cho một phần tử dữ liệu X, bộ phân loại lớp sẽ gán nhãn cho X là lớp có xác suất hậu nghiệm lớn nhất. Cụ thể, bộ phân loại Bayes sẽ dự đoán X thuộc vào lớp  $C_i$  nếu và chỉ nếu:

$$P(C_i|X) > P(C_j|X) \quad (1 \leq i \leq m, i \neq j) \quad (2.8)$$

Giá trị này sẽ được tính dựa vào định lý Bayes:  $P(C_i|X) = \frac{p(X|C_i) \times P(C_i)}{P(X)}$

3. Để tìm giá trị xác suất lớn nhất, ta nhận thấy trong công thức (2.8) giá trị  $P(X)$  là giống nhau với mọi lớp nên không cần tính. Do đó chỉ cần tìm giá trị lớn nhất của  $P(X|C_i) \times P(C_i)$ . Trong đó,  $P(C_i)$  được ước lượng bằng công thức  $P(C_i) = \frac{|D_i|}{|D|}$  với  $D_i$  là tập các phần tử dữ liệu thuộc lớp  $C_i$ . Nếu xác suất tiên nghiệm  $P(C_i)$  cũng không xác định được thì ta coi chúng bằng nhau, khi đó chỉ cần tìm giá trị  $P(X|C_i)$  lớn nhất.
4. Khi số lượng các thuộc tính mô tả dữ liệu là lớn thì chi phí tính toán  $P(X|C_i)$  là rất lớn, do đó để làm giảm độ phức tạp, giải thuật Naïve Bayes giả thiết các thuộc tính là độc lập nhau hay không có sự phụ thuộc nào giữa các thuộc tính. Khi đó ta có:

$$P(X|C_i) = \prod_{k=1}^n P(x_k|C_i) = P(x_1|C_i) \dots \times P(x_n|C_i) \quad (2.9)$$

Naïve Bayes là một giải thuật đơn giản, dễ cài đặt, thời gian huấn luyện nhanh, thực hiện phân loại khá tốt với các bài toán đa nhãn và không cần quá nhiều dữ liệu huấn luyện. Tuy nhiên, giả định về sự độc lập giữa các đặc trưng của dữ liệu thường khó xảy ra trong thế giới thực. Với những đặc điểm trên, Naïve Bayes thường được sử dụng trong các hệ thống dự đoán thời gian thực, các bài toán dự đoán đa nhãn, phân loại văn bản, lọc thư rác, ...

#### 2.3.3.2. Support Vector Machine (SVM)

Support Vector Machines (SVM) là kỹ thuật học có giám sát được đề xuất lần đầu tiên vào năm 1992 cho bài toán phân loại nhị phân. Hiện nay thuật toán này được mở rộng cho các bài toán phân loại đa lớp.

SVM hỗ trợ xây dựng một siêu phẳng hoặc một tập hợp các siêu phẳng trong một không gian nhiều chiều hoặc vô hạn chiều, có thể được sử dụng cho phân loại, hồi quy

hoặc các nhiệm vụ khác. Để phân loại tốt nhất thì các siêu phẳng nằm càng xa các điểm dữ liệu của tất cả các lớp (gọi là lề) càng tốt. Trong nhiều trường hợp, không thể phân chia các lớp dữ liệu một cách tuyến tính trong một không gian ban đầu vì vậy, cần phải ánh xạ các điểm dữ liệu trong không gian ban đầu vào một không gian mới nhiều chiều hơn, để việc phân tách chúng trở nên dễ dàng hơn.

Để việc tính toán được hiệu quả, phép ánh xạ sử dụng trong thuật toán SVM chỉ ràng buộc tích vô hướng của các véc-tơ dữ liệu trong không gian mới có thể được tính dễ dàng từ các tọa độ trong không gian cũ.

$$K(a, b) = \langle a, b \rangle \quad (2.10)$$

Sử dụng hàm đối ngẫu Lagrange, bài toán tìm lế cực đại của SVM được đưa về bài toán tìm véc-tơ hệ số  $\vec{\alpha} = (\alpha_1, \dots, \alpha_n)$  cho phép cực tiểu hóa hàm mục tiêu

$$W(\vec{\alpha}) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i y_j \alpha_i \alpha_j K(x_i, x_j) \quad (2.11)$$

đồng thời thỏa mãn:

$$\sum_{i=1}^n y_i \alpha_i = 0, 0 \leq \alpha_i \leq C \quad (2.12)$$

Trong đó,  $x_i$  và  $y_i$  tương ứng là dữ liệu và nhãn của nó,  $\alpha_i$  là hệ số cần xác định,  $C$  là số điểm dữ liệu tối đa được phân loại sai.

Quá trình huấn luyện SVM là quá trình xác định  $\alpha_i$ . Phương pháp hiệu quả và thông dụng nhất là tối ưu tuần tự SMO [10]. Sau khi phân loại xong, giá trị nhãn phân loại cho mẫu mới được tính bởi:

$$f(x) = \text{sign} \left( \sum_{i=1}^n y_i \alpha_i K(x_i, x) + b \right) \quad (2.13)$$

với  $b$  được tính trong giai đoạn huấn luyện theo công thức:

$$b = y_i - \sum_{j=1}^n y_j \alpha_j K(x_i, x_j) \quad (2.14)$$

Trong đó,  $i$  là một hệ số thỏa mãn điều kiện  $0 < \alpha_i < C$ .

### 2.3.3.3. Logistic Regression

Logistic Regression có công thức biểu diễn như sau:

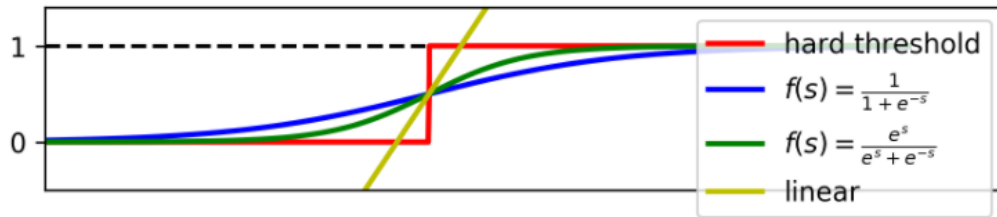
$$f(x) = \theta(w^T x) \quad (2.15)$$

Trong đó,  $f(x)$  là xác suất sinh viên đỗ hay trượt,  $x$  là số giờ sinh viên ôn tập,  $w^T$  là hằng số được huấn luyện sao cho kết quả dự đoán là chính xác nhất,  $\theta$  là hàm kích hoạt

đưa kết quả về dạng xác suất. Tuy nhiên, các bài toán trong thực tế thường có dữ liệu có nhiều đặc trưng, cho nên  $x = (x_1, x_2, \dots, x_n)$  là một véc-tơ,  $w$  là ma trận các hằng số.

Một số hàm kích hoạt cho mô hình tuyến tính được mô tả trong Hình 2.3. Đường màu đỏ và vàng không phù hợp với bài toán. Đường màu vàng không bị chặn ở 2 đầu. Ngoài ra, các điểm dữ liệu trong bài toán không hoàn toàn phân tách nên đường màu đỏ không phù hợp. Các đường màu xanh lam và xanh lục phù hợp với bài toán của đã nêu hơn. Chúng có một vài tính chất quan trọng sau:

- Là hàm số liên tục nhận giá trị thực, bị chặn trong khoảng  $(0, 1)$
- Nếu coi điểm có tung độ là  $1/2$  làm điểm phân chia thì các điểm càng xa điểm này về phía bên trái có giá trị càng gần 0. Ngược lại, các điểm càng xa điểm này về phía phải có giá trị càng gần 1. Điều này khớp với nhận xét rằng học càng nhiều thì xác suất đổ càng cao và ngược lại.
- Mượt (smooth) nên có đạo hàm mọi nơi, có thể được lợi trong việc tối ưu.



Hình 2.3: Các hàm kích hoạt

### Hàm phi tuyến

Hàm Sigmoid rất hay được sử dụng vì nó bị chặn trong khoảng  $(0, 1)$ . Hơn nữa, đạo hàm của hàm sigmoid rất đơn giản nên nó được sử dụng rộng rãi.

$$f(x) = \frac{1}{1 + e^{-x}} \triangleq \sigma(x) \quad (2.16)$$

$$\sigma'(x) = \frac{e^{-x}}{(1 + e^{-x})^2} = \frac{1}{1 + e^{-x}} \frac{e^{-x}}{1 + e^{-x}} = \sigma(x)(1 - \sigma(x)) \quad (2.17)$$

Ngoài ra, hàm **tanh** cũng hay được sử dụng. Hàm số này nhận giá trị trong khoảng  $(-1, 1)$ .

$$\tanh(s) = \frac{e^s - e^{-s}}{e^s + e^{-s}} \quad (2.18)$$

### Hàm mất mát và phương pháp tối ưu

Công thức cập nhật cho Logistic Regression sử dụng hàm Sigmoid theo phương pháp Stochastic Gradient Descent với điểm dữ liệu  $(x_i, y_i)$  là:

$$w = w + \mu(y_i - z_i)x_i \quad (2.19)$$

Trong đó,  $z_i = \theta(w^T x_i)$ .

Tuy có tên là Regression, nhưng thuật toán này thường sử dụng nhiều trong các bài toán phân loại. Mô hình này phân loại dữ liệu dựa trên phương trình siêu phẳng  $w^T x$  có dạng tuyến tính. Do vậy, mô hình này chỉ phù hợp với các loại dữ liệu mà 2 lớp là phân biệt tuyến tính. Logistic Regression không phù hợp với các loại dữ liệu có lớp nằm bên ngoài đường tròn, lớp nằm trong đường tròn đó. Ngoài ra, các điểm dữ liệu nhiều sẽ ảnh hưởng rất nhiều tới độ chính xác của mô hình.

## 2.4. KẾT LUẬN

Nội dung Chương 2 đã tập trung trình bày về mô hình kết hợp hành vi đánh giá và kết hợp cùng với đó là cách triển khai mô hình này vào hệ tư vấn. Ngoài ra, nội dung chương còn trình bày về các kỹ thuật tiền xử lý dữ liệu văn bản và các mô hình phân loại được đồ án sử dụng. Với kỳ vọng cải thiện độ chính xác kết quả của hệ tư vấn, nội dung Chương 3 sẽ trình bày về các kết quả thử nghiệm và đánh giá của mô hình này cùng với bộ dữ liệu mà đồ án sử dụng.



## CHƯƠNG 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ

Nội dung của chương này tập trung trình bày về các kết quả thử nghiệm để đánh giá hiệu quả dự đoán của mô hình được trình bày trong Chương 2. Ngoài ra, bộ dữ liệu được sử dụng cho thử nghiệm cũng được trình bày trong chương này. Nội dung của chương này gồm 7 phần:

- Phát biểu bài toán
- Dữ liệu thử nghiệm
- Môi trường thử nghiệm
- Tiêu chí đánh giá
- Kết quả thử nghiệm
- Tổng hợp đánh giá
- Kết luận

### 3.1. PHÁT BIỂU BÀI TOÁN

Với kỳ vọng mô hình kết hợp đã được trình bày trong Chương 2 đồ án mang lại kết quả tốt, đồ án tiến hành so sánh mô hình này với mô hình dữ liệu hành vi đánh giá. Tuy nhiên, trước khi tiến hành đánh giá mô hình kết hợp, đồ án cần khảo sát các mô hình phân loại văn bản và thuật toán lọc cộng tác để tìm ra 2 mô hình tốt nhất, sử dụng cho bài toán kết hợp. Các mô hình phân loại văn bản được đồ án thực hiện khảo sát: Logistic Regression, Naïve Bayes, SVM. Với thuật toán lọc cộng tác, đồ án thực hiện khảo sát: Matrix Factorization (MF) và Item-item Collaborative Filtering (II-CF). Như vậy, đồ án cần giải quyết 4 bài toán được trình bày ngay sau đây.

**Bài toán 1: Khảo sát các mô hình phân loại văn bản.** Tại đây, đồ án tiến hành đánh giá kết quả dự đoán của các mô hình phân loại văn bản được khảo sát: Naïve Bayes, SVM, Logistic Regression khi sử dụng TF-IDF với N-grams khác nhau. Từ đó chọn ra mô hình phân loại cùng với giá trị N-grams cho kết quả tốt nhất trong số các mô hình được khảo sát.

**Bài toán 2: Khảo sát các thuật toán lọc cộng tác.** Trong bài toán này, đồ án tiến hành đánh giá các mô hình dự đoán được huấn luyện bởi các thuật toán tư vấn: MF và II-CF. Dữ liệu đầu vào của thuật toán là các hành vi đánh giá của người dùng dành cho các khách sạn. Bài toán được thực hiện nhằm mục đích chọn ra thuật toán tư vấn cho ra mô hình tốt nhất trong các thuật toán được khảo sát.

**Bài toán 3: Khảo sát các giá trị  $\alpha$  và  $\beta$  trong công thức (2.3).** Việc thay đổi 2 tham số này nhằm xác định:

- Đánh giá, chứng minh mô hình kết hợp sẽ nâng cao hiệu quả tư vấn.
- Tìm giá trị tham số  $\alpha$  và  $\beta$  cho hiệu quả dự đoán tốt nhất để áp dụng cho việc kết hợp giữa dữ liệu hành vi đánh giá và dữ liệu hành vi bình luận.



Sau khi thực hiện Bài toán 1 và 2, đồ án xác định được mô hình phân loại văn bản và mô hình tư vấn tốt nhất trong số các mô hình đã thử nghiệm. Hai mô hình này sẽ được chọn để thực hiện đánh giá mô hình kết hợp.

### 3.2. DỮ LIỆU THỬ NGHIỆM

Để giải quyết các bài toán đã đề ra, sinh viên sử dụng 2 bộ dữ liệu cho việc xây dựng hệ thống. Bộ dữ liệu thứ nhất, Booking được dùng cho bài toán phân loại quan điểm tiếng Anh của người dùng. Bộ dữ liệu thứ hai, Tripadvisor được dùng cho các bài toán dự đoán điểm đánh giá. Chi tiết về 2 bộ dữ liệu được trình bày ngay sau đây.

#### 3.2.1. Bộ dữ liệu Booking

##### 3.2.1.1. Giới thiệu

Bộ dữ liệu Booking<sup>1</sup> là bộ dữ liệu được sinh viên sử dụng cho Bài toán 1, được thu thập từ trang web Booking.com. Trong bộ dữ liệu chứa các trường thông tin:

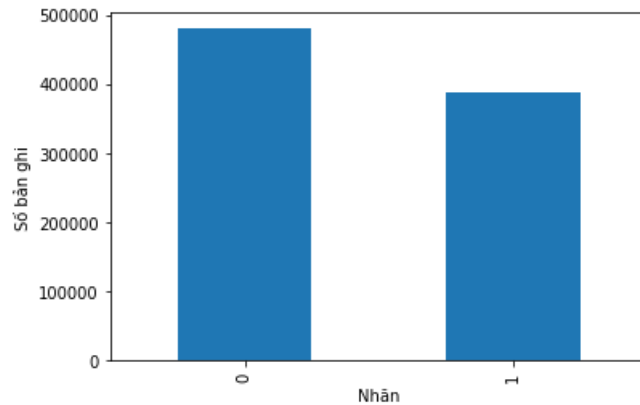
- Hotel\_Address: Địa chỉ khách sạn
- Review\_Date: Ngày người dùng thực hiện đánh giá
- Average\_Score: Điểm đánh giá trung bình của khách sạn
- Hotel\_Name: Tên của khách sạn
- Reviewer\_Nationality: Quốc tịch của người đánh giá
- Negative\_Review: Phần bình luận du khách tỏ ra không thích về khách sạn trong toàn bộ đánh giá của mình. Nếu không có phần nào tiêu cực, giá trị của mục này là “No Negative”.
- Review\_Total\_Negative\_Word\_Counts: Tổng số từ trong phần bình luận tiêu cực.
- Positive\_Review: Phần bình luận du khách tỏ ra thích về khách sạn trong toàn bộ đánh giá của mình. Nếu không có phần nào tích cực, giá trị của mục này là “No Positive”.
- Review\_Total\_Positive\_Word\_Counts: Tổng số từ trong phần bình luận tích cực.
- Reviewer\_Score: Điểm đánh giá người dùng dành cho khách sạn.
- Total\_Number\_of\_Reviews\_Reviewer\_Has\_Given: Tổng số đánh giá mà người dùng đã thực hiện trong quá khứ.
- Total\_Number\_of\_Reviews: Tổng số đánh giá mà khách sạn có.
- Tags: Nhãn mà khách sạn được gán bởi người dùng.
- Lat: Vĩ độ của khách sạn.
- Lng: Kinh độ của khách sạn.

Trong mỗi bình luận của người dùng thường có cả 2 phần tiêu cực và tích cực. Nếu trực tiếp sử dụng những bình luận từ người dùng thì sẽ ảnh hưởng tới độ chính xác của mô hình. Tuy nhiên, bộ dữ liệu đã tách riêng 2 phần tích cực và tiêu cực của bình luận, do đó đồ án không cần tiền xử lý các dữ liệu bình luận này. Các phần bình luận tích cực

---

<sup>1</sup> <https://www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe>

và tiêu cực trong bộ dữ liệu lần lượt là 1.031.476 và 867.640. Hình 3.1 cho thấy sự phân bố dữ liệu khi thực hiện tách các phần tích cực, tiêu cực từ bình luận ban đầu khá cân bằng.



Hình 3.1: Sự phân bố của các phần bình luận

### 3.2.2. Tripadvisor Dataset

#### 3.2.2.1. Giới thiệu

Mặc dù chứa rất nhiều thông tin, bộ dữ liệu lại không chứa thông tin về người dùng thực hiện đánh giá. Điều này khiến bộ dữ liệu không phù hợp cho giải quyết Bài toán 2 và 3. Vì lý do trên, sinh viên chỉ sử dụng thông tin về những bình luận của người dùng đã được gán nhãn: tích cực, tiêu cực, trong bộ dữ liệu để phục vụ cho Bài toán 1. Đây cũng là lý do bộ dữ liệu Tripadvisor Dataset được sử dụng trong đồ án này.

Tripadvisor Dataset<sup>2</sup> là bộ dữ liệu sinh viên sử dụng cho Bài toán 2 và 3, được thu tập từ trang web Tripadvisor.com, thực hiện bởi tác giả Myle Ott<sup>3</sup>, một nhà nghiên cứu về xử lý ngôn ngữ tự nhiên làm việc tại Facebook AI Research. Bộ dữ liệu gồm 2 phần: phần thông tin khách sạn và phần thông tin đánh giá.

Phần thông tin khách sạn chứa dữ liệu về 4333 khách sạn khác nhau dưới dạng Json. Mỗi khách sạn có những trường thông tin sau:

- id: ID của khách sạn
- name: Tên khách sạn
- hotel\_class: Hạng khách sạn
- region\_id: ID vùng
- url: liên kết tới trang thông tin khách sạn trên trang Tripadvisor
- phone: Số điện thoại khách sạn
- details: Thông tin chi tiết về khách sạn
- address: Mục này chứa các thông tin về địa chỉ khách sạn
  - street-address: Địa chỉ cụ thể
  - postal-code: Mã bưu điện

<sup>2</sup> <https://www.cs.cmu.edu/~jiwei1/html/hotel-review.html>

<sup>3</sup> <https://myleott.com>

- locality: Tên thành phố

```
{'address': {'locality': 'New York City',
'postal-code': '10036',
'region': 'NY',
'street-address': '147 West 43rd Street'},
'details': None,
'hotel_class': 4.0,
'id': 113317,
'name': 'Casablanca Hotel Times Square',
'phone': '',
'region_id': 60763,
'type': 'hotel',
'url': 'http://www.tripadvisor.com/Hotel_Review-g60763-d113317-Reviews-Casablanca_Hotel_Times_Square-New_York_City_New_York.html'}
```

Hình 3.2: Một bản ghi trong phần thông tin khách sạn

Phần thông tin đánh giá khách sạn chứa 878561 đánh giá từ người dùng dưới dạng Json. Mỗi dữ liệu đánh giá có những trường thông tin sau:

- id: ID của bản ghi
- author: Trong mục này chứa các trường thông tin liên quan đến người thực hiện đánh giá
  - id: ID của người dùng
  - location: vị trí gửi đánh giá
  - num\_helpful\_votes: Số lượng bầu chọn từ người dùng khác cho rằng đánh giá là có ích
  - num\_reviews: Số lượt người dùng đã đánh giá
  - username: Tên người dùng
- date: Ngày thực hiện đánh giá
- date\_stayed: Ngày du khách ở khách sạn
- offering\_id: ID khách sạn được đánh giá bởi người dùng
- title: Tiêu đề đánh giá
- text: Bình luận của người dùng
- ratings: Trong mục này chứa các thông tin liên quan đến các hạng mục đánh giá. Điểm đánh giá trong khoảng từ 0 đến 5.
  - cleanliness: Điểm đánh giá sự sạch sẽ
  - location: Điểm đánh giá cho vị trí khách sạn
  - rooms: Điểm đánh giá về phòng của khách sạn
  - service: Điểm đánh giá về dịch vụ
  - sleep\_quality: Điểm đánh giá về chất lượng ngủ
  - value: Điểm đánh giá về giá trị trải nghiệm
  - overall: Điểm đánh giá tổng quan

```
{
  'author': {
    'id': '8C0B42FF3C0FA366A21CFD785302A032',
    'location': 'Gold Coast',
    'num_cities': 22,
    'num_helpful_votes': 12,
    'num_reviews': 29,
    'num_type_reviews': 24,
    'username': 'Papa_Panda'},
  'date': 'December 17, 2012',
  'date_stayed': 'December 2012',
  'id': 147643103,
  'num_helpful_votes': 0,
  'offering_id': 93338,
  'ratings': {
    'cleanliness': 5.0,
    'location': 5.0,
    'overall': 5.0,
    'rooms': 5.0,
    'service': 5.0,
    'sleep_quality': 5.0,
    'value': 5.0},
  'text': 'Stayed in a king suite for 11 nights and yes it cots us a bit but we were happy with the standard of room, the location and the friendl',
  'title': '"Truly is "Jewel of the Upper Wets Side"',
  'via_mobile': False}
}
```

Hình 3.3: Một bản ghi trong phần thông tin đánh giá từ người dùng

### 3.2.2.2. Phân chia bộ dữ liệu

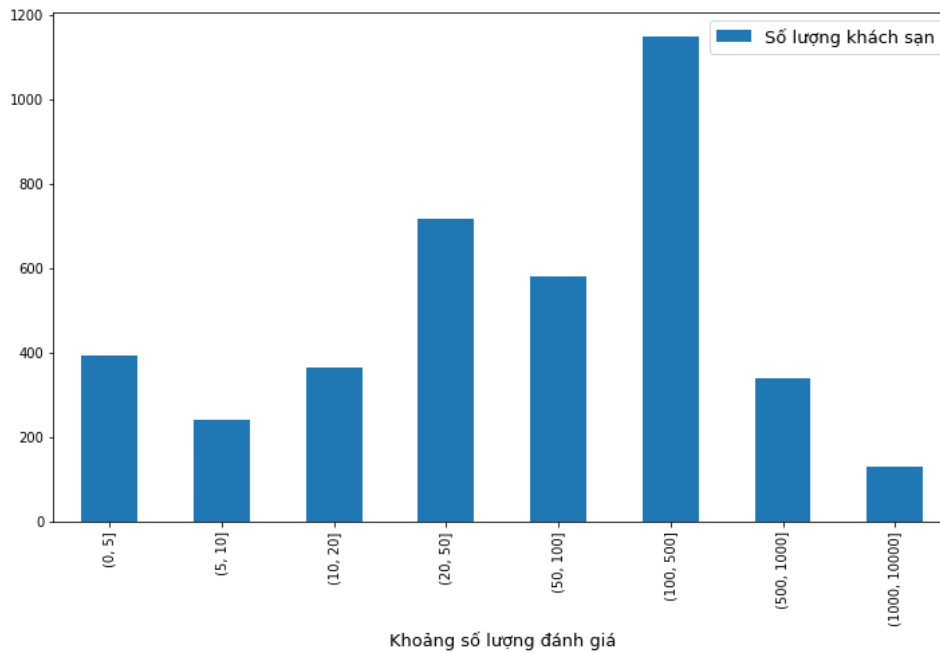
Trong bộ dữ liệu, một số bản ghi không có ID của người dùng thực hiện đánh giá, một thông tin quan trọng để thực hiện Bài toán 2 và 3. Do đó, sinh viên thực hiện loại bỏ các bản ghi này và bộ dữ liệu còn lại có kích thước 801.196. Bảng 3.1 thống kê số số lượng người dùng đưa ra bao nhiêu đánh giá. Trong đó, 456.989 tương ứng với 57,04% số lượng người dùng chỉ đưa ra 1 đánh giá cho khách sạn.

Hình 3.4 mô tả thống kê số lượng đánh giá cho khách sạn theo khoảng. Biểu đồ cho thấy, hầu hết khách sạn đều được đánh giá từ 20 đến 500 lượt. Hình 3.5 là phân bố điểm đánh giá được đưa ra bởi người dùng. Có thể thấy, các điểm đánh giá 4 và 5 chiếm đa số trong bộ dữ liệu.

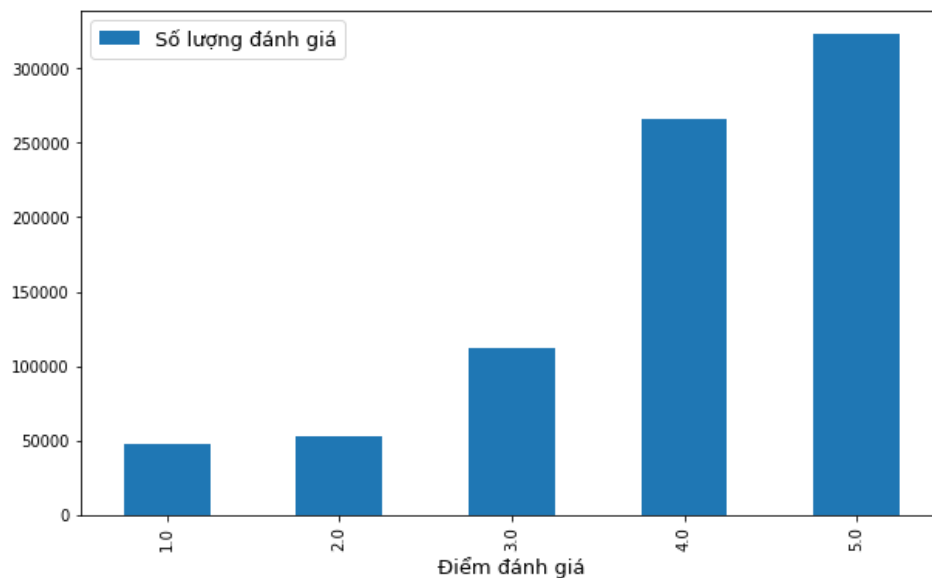
Bảng 3.1: Thống kê số số người dùng theo số lượng đánh giá được đưa ra

STT	Số đánh giá được đưa ra	Số người dùng
1	1	456.989
2	2	73.588
3	3	24.126
4	4	10.012
5	5	4.720
6	6	2.624
7	7	1.428
8	8	951
9	9	616
10	10	439
11	11	304
12	12	197
13	13	153
14	>13	540
Tổng		576687

Thống kê trong bảng Bảng 3.1 cho thấy số lượng người dùng chỉ đưa ra 1 và 2 đánh giá chiếm 75,41% số bản ghi trong bộ dữ liệu. Điều này gây ra hiện tượng thừa thớt dữ liệu trong ma trận tiện ích, làm giảm độ chính xác của dự đoán.



Hình 3.4: Thống kê phân bố số lượng đánh giá cho khách sạn theo khoảng



Hình 3.5: Phân bố điểm đánh giá

Để khắc phục hiện tượng này, đồ án thực hiện thí nghiệm trên các bộ dữ liệu con. Mỗi bộ dữ liệu con này có được bằng cách kết hợp 2 bộ dữ liệu được lọc từ bộ dữ liệu ban đầu. Ví dụ, bộ dữ liệu ds\_11\_2310 được tạo thành khi ghép 2 bộ dữ liệu có kích thước tương đương, bộ thứ nhất thỏa mãn điều kiện mỗi người dùng đánh giá ít nhất 11

lần và bộ thứ hai thỏa mãn mỗi khách sạn trong đó đều được đánh giá ít nhất 2310 lần. Đồ án tiến hành thực nghiệm với các bộ dữ liệu được liệt kê trong bảng Bảng 3.2.

Bảng 3.2: Bảng phân chia các bộ dữ liệu theo các tiêu chí khác nhau

Tên bộ dữ liệu	Kích thước
ds_4_1200	229.895
ds_8_2100	69.075
ds_11_2310	34.765
df_14_2430	20.273
ds_15_2500	16.157
ds_17_2570	11.281

### 3.3. MÔI TRƯỜNG THỬ NGHIỆM

Đồ án sử dụng Python với IDE Pycharm để thực hiện cài đặt và khảo sát.

Bảng 3.3: Môi trường thử nghiệm

Thành phần	Thông số
Hệ điều hành	Windows 10 Pro 64bit
Bộ vi xử lý	I5-7300HQ 2,5Hz
Bộ nhớ trong	8Gb
Bộ nhớ ngoài	1Tb

Bảng 3.4: Thư viện hỗ trợ chính

Tên thư viện	Chức năng
pandas	hỗ trợ xử lý dữ liệu đầu vào, xuất tệp dữ liệu
numpy	hỗ trợ tính toán ma trận
sklearn	hỗ trợ cài đặt các mô hình phân loại văn bản

### 3.4. TIÊU CHÍ ĐÁNH GIÁ

Để đánh giá hiệu năng, mô hình dự đoán và tư vấn cần có những tiêu chí định lượng cụ thể để so sánh và đánh giá. Trong phạm vi đồ án, các mô hình phân loại văn bản được đánh giá dựa trên 2 tiêu chí: F1-Score và Accuracy, các mô hình tư vấn được đánh giá dựa trên 2 tiêu chí: RMSE và MAE.

#### 3.4.1. Tiêu chí đánh giá sử dụng cho phân loại văn bản

Trong lĩnh vực máy học và cụ thể là vấn đề phân loại thống kê, ma trận hỗn độn hay còn được gọi là ma trận lỗi, là một bảng cụ thể cho phép người dùng hình dung hiệu suất của một thuật toán, thường là một thuật toán được giám sát. Mỗi hàng của ma trận đại diện cho các trường hợp trong một nhãn thực tế. Trong khi đó, mỗi cột của ma trận đại diện cho 1 trường hợp có thể được thuật toán dự đoán.

Bảng 3.5: Bảng ma trận hỗn độn

Nhãn của dữ liệu	Dự đoán của mô hình	
	Tích cực (0)	Tiêu cực (1)
Tích cực (0)	True Positive (TP)	False Negative (FN)
Tiêu cực (1)	False Positive (FP)	True Negative (TN)

Bảng 3.5 là ma trận hỗn độn khi thực hiện bài toán phân loại quan điểm từ bình luận người dùng. Trong đó, nhãn “0” là nhãn mang ý nghĩa tích cực, nghĩa là trong bình luận của người dùng bày tỏ ý kiến thích khách sạn. Ngược lại, với nhãn “1”, người dùng để lại bình luận mang nhãn này bày tỏ ý kiến không thích. Ma trận gồm 4 ô: TP, FN, FP, TN.

- True Positive (TP) là số dự đoán nhãn “0” của hệ thống đúng so với nhãn thực.
- False Positive (FP) là số dự đoán nhãn “0” của hệ thống nhưng sai so với nhãn thực.
- True Negative (TN) là số dự đoán nhãn “1” của hệ thống đúng với nhãn thực.
- False Negative (FN) là số dự đoán nhãn “1” của hệ thống nhưng sai với nhãn thực.

Từ ma trận hỗn độn, ta có 4 tiêu chí đánh giá: Accuracy, Precision, Recall, F-Score.

1. Accuracy là tỉ lệ chính xác trong các dự đoán của thuật toán, được tính bằng cách lấy tỉ lệ của các dự đoán đúng  $N_{true}$  so với tổng các dự đoán  $N$ . Accuracy được tính theo công thức sau:

$$Accuracy = \frac{N_{true}}{N} = \frac{TP + TN}{TP + TN + FP + FN}$$

2. Precision là tỉ lệ dự đoán số lượng nhãn “0” đúng so với số lượng nhãn “0” mà thuật toán dự đoán.

$$Precision = \frac{TP}{TP + FP}$$

3. Recall là tỉ lệ của số dự đoán nhãn “0” đúng so với số lượng nhãn “0” thực tế.

$$Recall = \frac{TP}{TP + FN}$$

4. F1-Score là tiêu chí trung hòa giữa Precision và Recall

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

### 3.4.2. Tiêu chí đánh giá sử dụng cho hệ tư vấn lọc cộng tác

Trong đồ án, sinh viên chọn 2 độ đo RMSE và MAE để đánh giá hiệu năng của hệ tư vấn [11].

RMSE (Root Mean Square Error) là căn của sai số bình phương trung bình, có được bằng cách lấy căn bậc 2 của trung bình cộng bình phương hiệu giữa tất cả các cặp điểm đánh giá thật và dự đoán [12]. Công thức tổng quát của RMSE được định nghĩa như sau:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (p_i - r_i)^2}$$

MAE (Mean Absolute Error) là sai số tuyệt đối trung bình, được tính bằng cách lấy trung bình cộng của tất cả giá trị tuyệt đối giữa điểm đánh giá thật và điểm đánh giá dự đoán [12]. Công thức tổng quát của MAE được định nghĩa như sau:

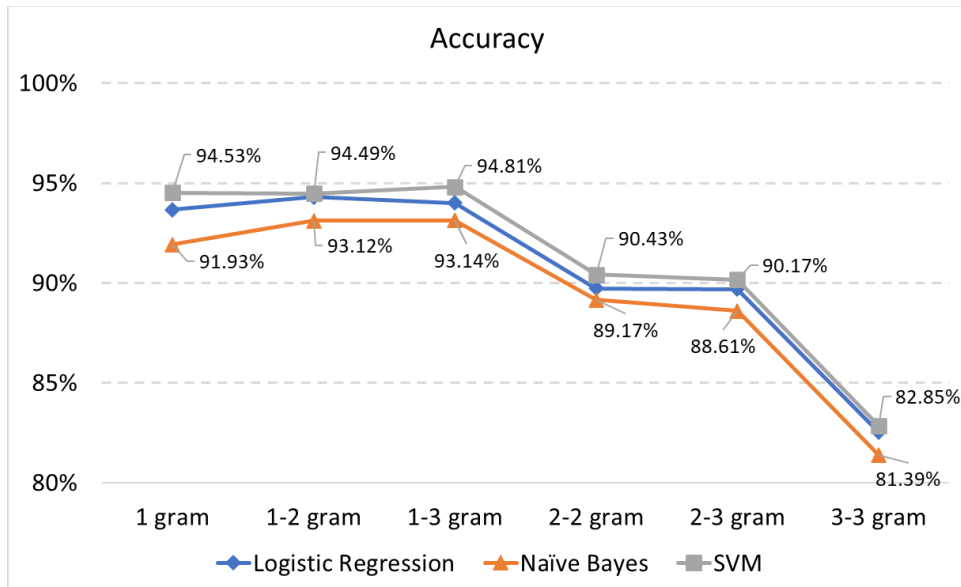
$$MAE = \frac{1}{n} \sum_{i=1}^n |p_i - r_i|$$

Trong 2 công thức trên,  $p_i$  là điểm đánh giá được thuật toán dự đoán,  $r_i$  là điểm đánh giá thực tế.

### 3.5. KẾT QUẢ THỬ NGHIỆM

#### 3.5.1. Bài toán 1: Khảo sát các mô hình phân loại văn bản

Hình 3.6 là biểu đồ so sánh Accuracy của 3 mô hình phân loại khi tham số N-grams thay đổi. Có thể thấy mô hình SVM với tham số N-gram là khoảng từ 1 tới 3 cho kết quả cao nhất với Accuracy là 94,81%. Mô hình Naïve Bayes cho kết quả kém nhất trong 3 mô hình được thử nghiệm.

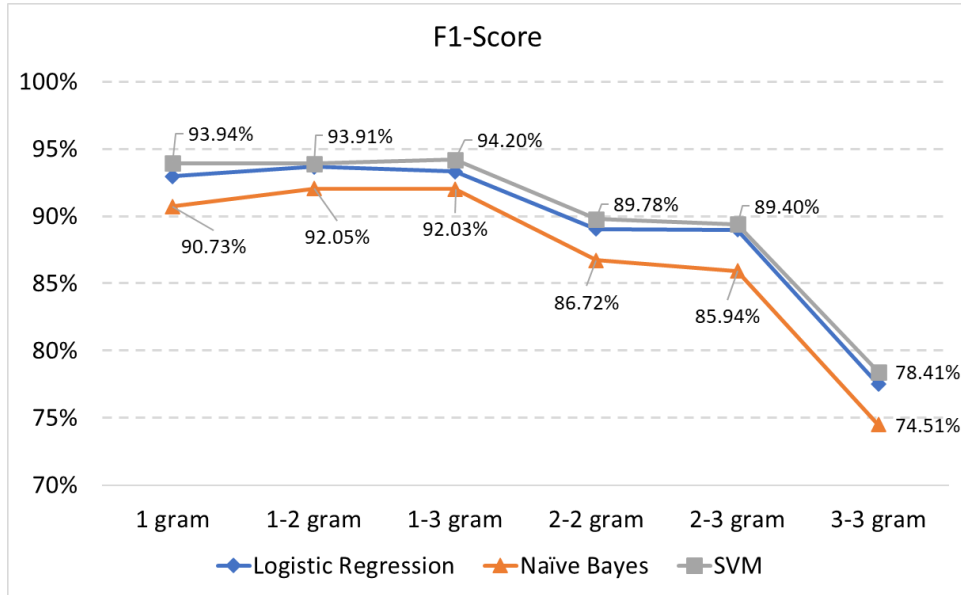


Hình 3.6: Biểu đồ so sánh Accuracy giữa 3 mô hình phân loại

Hình 3.7 là biểu đồ so sánh F1-Score giữa 3 mô hình phân loại khi tham số N-grams thay đổi. Mô hình SVM cho kết quả tốt nhất với mọi tham số N-grams. Cao hơn cả là



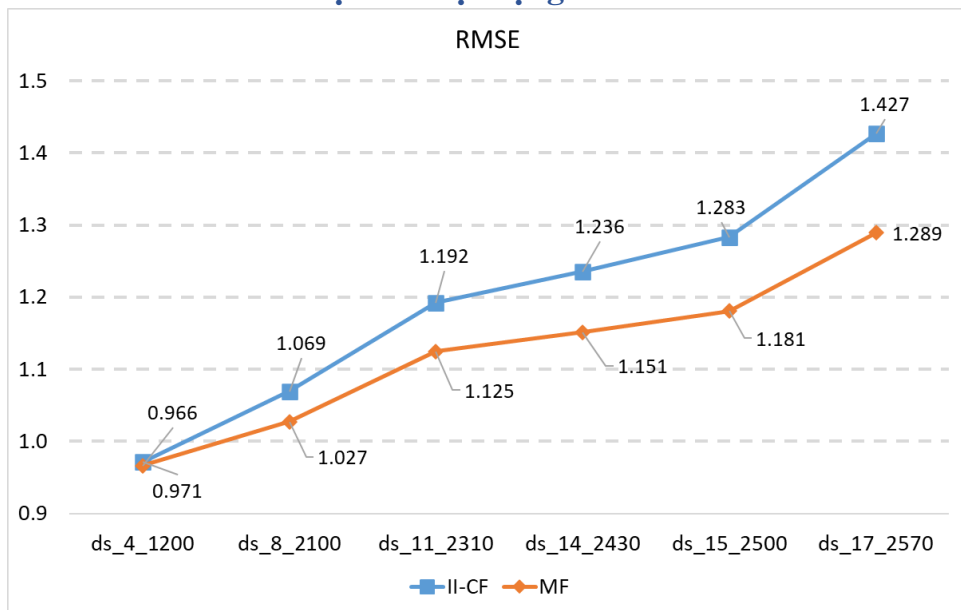
N-grams trong khoảng 1-3, SVM có F1-Score 94,2%. Naïve Bayes là mô hình kém nhất trong 3 mô hình được thử nghiệm.



Hình 3.7: Biểu đồ so sánh F1-Score giữa 3 mô hình phân loại

Naïve Bayes cho kết quả kém nhất trong 3 mô hình ở cả 2 chỉ số có thể là do thuật toán này hoạt động dựa trên nguyên lý độc lập xác suất. Naïve Bayes sẽ coi các đặc trưng của dữ liệu đầu vào là độc lập với nhau. Nhưng với dữ liệu văn bản, các từ trong câu và các câu trong đoạn văn có ý nghĩa liên quan tới nhau. Điều này dẫn tới kết quả thấp của mô hình này.

### 3.5.2. Bài toán 2: Khảo sát thuật toán lọc cộng tác

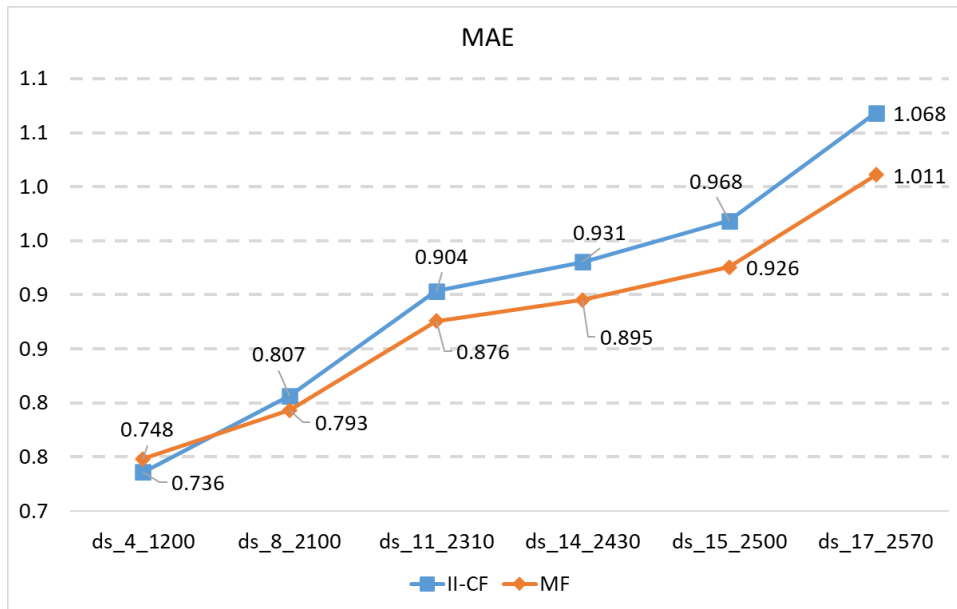


Hình 3.8: Biểu đồ so sánh MF-CF và II-CF theo tiêu chí RMSE

Biểu đồ so sánh RMSE của 2 thuật toán II-CF và MF trên các bộ dữ liệu có mật độ đánh giá tăng dần, thể hiện trong Hình 3.8. Dễ dàng nhận thấy, chỉ số này tăng dần khi

số lượng dữ liệu huấn luyện giảm dần. Ngoài ra, MF tốt hơn IICF ở hầu hết các bộ dữ liệu. Khi mật độ đánh giá càng dày đặc thì RMSE của MF càng bỏ xa II-CF. Tại bộ dữ liệu có mật độ đánh giá dày đặc nhất được thử nghiệm, ds\_17\_2570, RMSE của MF tốt hơn khoảng 10% so với II-CF.

Hình 3.9 là biểu đồ so sánh MAE của 2 thuật toán với các bộ dữ liệu có mật độ đánh giá tăng dần. Điều tương tự cũng xảy ra với tiêu chí đánh giá MAE khi tăng dần theo chiều giảm của số lượng dữ liệu huấn luyện. Ngoài ra, cũng giống với RMSE, khi mật độ dữ liệu càng dày đặc thì MAE của MF càng bỏ xa II-CF. Tại bộ dữ liệu ds\_17\_2570, MF có chỉ số MAE tốt hơn II-CF 5,33%. Tuy nhiên, tại bộ dữ liệu có mật độ thưa thớt nhất, ds\_4\_1200 chỉ số MAE của II-CF lại tốt hơn 1,6% so với MF.

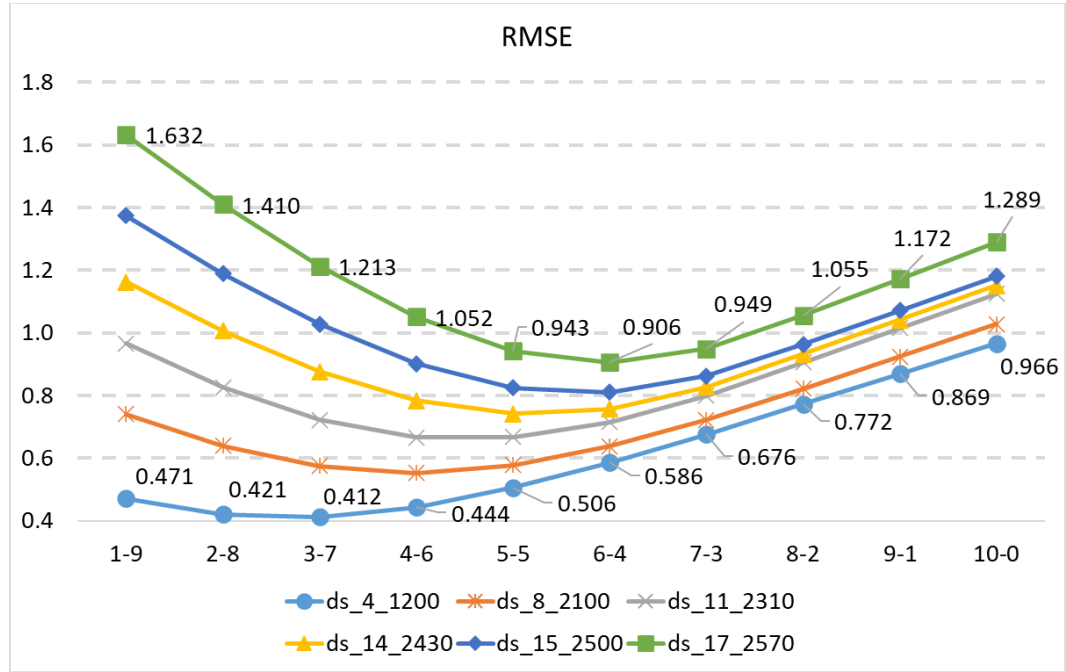


Hình 3.9: Biểu đồ so sánh MF-CF và II-CF theo tiêu chí MAE

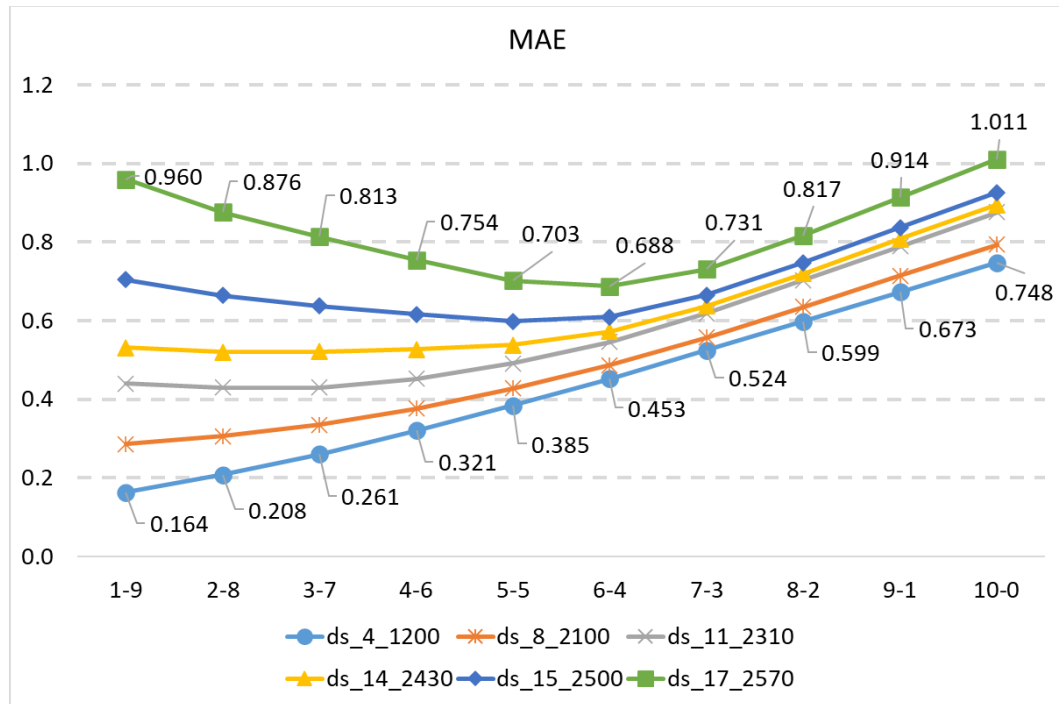
### 3.5.3. Bài toán 3: Khảo sát các giá trị $\alpha$ và $\beta$ trong công thức (2.3)

Sau khi giải quyết Bài toán 1 và 2, mô hình phân loại văn bản tốt nhất trong số các mô hình được thử nghiệm là SVM và thuật toán tư vấn tốt nhất trong số các thuật toán được thử nghiệm là MF. Vì vậy, đề án lựa chọn SVM và MF để thực hiện đánh giá mô hình kết hợp.

Hình 3.10 là biểu đồ so sánh RMSE khi thay đổi tỷ lệ  $\alpha$  và  $\beta$  trong công thức (2.3) được thử nghiệm trên các bộ dữ liệu có mật độ đánh giá tăng dần. Dễ dàng nhận thấy với các bộ dữ liệu có mật độ đánh giá thấp, tỷ lệ  $\alpha$ - $\beta$  với  $\alpha$  thấp,  $\beta$  sẽ cho RMSE thấp. Tuy nhiên, khi mật độ dữ liệu tăng dần, tỷ lệ này thay đổi theo chiều hướng  $\alpha$  tăng lên,  $\beta$  giảm. Tại 2 bộ dữ liệu có mật độ đánh giá cao nhất là ds\_15\_2500 và ds\_17\_2570, tỷ lệ này là 5-5 và 6-4. Ngoài ra, mô hình kết hợp với tỷ lệ thích hợp sẽ cho kết quả tốt hơn dữ liệu chỉ có hành vi đánh giá (10-0).



Hình 3.10: Biểu đồ so sánh RMSE khi thay đổi tỷ lệ  $\alpha$ - $\beta$  khi thực nghiệm MF với từng bộ dữ liệu



Hình 3.11: Biểu đồ so sánh MAE khi thay đổi tỷ lệ  $\alpha$ - $\beta$  khi thực nghiệm MF với từng bộ dữ liệu

Biểu đồ so sánh MAE khi thay đổi tỷ lệ  $\alpha$  và  $\beta$  trong công thức (2.3) được thể hiện trong Hình 3.11. Giống như RMSE, với các bộ dữ liệu có mật độ đánh giá thưa thớt, tỷ lệ  $\alpha$  và  $\beta$  với  $\alpha$  thấp,  $\beta$  cao sẽ cho MAE thấp, điển hình như ds\_4\_1200 là bộ dữ liệu có mật độ đánh giá thấp nhất. Tuy nhiên, khi mật độ đánh giá tăng lên, tỷ lệ này sẽ thay đổi theo chiều hướng  $\alpha$  tăng lên,  $\beta$  giảm xuống. Ngoài ra, có thể dễ dàng nhận thấy khi sử dụng

dùng mô hình kết hợp sẽ cho MAE thấp hơn khi chỉ sử dụng dữ liệu hành vi đánh giá (10-0).

### 3.6. TỔNG HỢP ĐÁNH GIÁ

Trong phần này, đồ án tổng hợp các kết quả quan trọng với những thử nghiệm đã thực hiện.

- **Bài toán 1: Khảo sát mô hình phân loại quan điểm người dùng:** SVM là mô hình cho kết quả phân loại tốt nhất với Accuracy và F1-Score lần lượt là 94,82% và 94,2%.
- **Bài toán 2: Khảo sát thuật toán lọc cộng tác:** Xét tổng thể, MF là thuật toán tư vấn tốt nhất trong các thuật toán được thử nghiệm. Ngoài ra, mật độ dữ liệu cũng ảnh hưởng tới độ chính xác của thuật toán. Cụ thể, mật độ dữ liệu đánh giá càng dày đặc thì MF càng cho kết quả tốt hơn II-CF. Tuy nhiên, với 1 trường hợp bộ dữ liệu thưa, II-CF cho RMSE sát xỉ MF, thậm chí có chỉ số MAE tốt hơn.
- **Đánh giá mô hình dữ liệu kết hợp:** Mô hình dữ liệu cho kết quả tốt hơn mô hình dữ liệu hành vi đánh giá khi kết hợp với tỷ lệ thích hợp. Khi mật độ dữ liệu thưa thớt, tỷ lệ  $\alpha$  và  $\beta$  sẽ có  $\alpha$  thấp,  $\beta$  cao. Tuy nhiên, khi mật độ dữ liệu trở nên dày đặc hơn thì tỷ lệ này dần tiến thay đổi theo chiều hướng  $\alpha$  tăng lên,  $\beta$  giảm xuống. Trong trường hợp ds\_17\_2570, bộ dữ liệu có mật độ đánh giá dày đặc nhất được thử nghiệm, tỷ lệ là 6-4 cho kết quả tốt nhất.

### 3.7. KẾT LUẬN

Trong Chương 3, đồ án đã đặt ra các bài toán cần được thử nghiệm, tập trung trình bày chi tiết các kết quả thử nghiệm. Trong phạm vi thực hiện thử nghiệm:

- SVM là mô hình phân loại văn bản tốt nhất
- MF là thuật toán tư vấn tốt nhất
- Mô hình kết hợp sẽ cho kết quả chính xác hơn khi  $\alpha$  và  $\beta$  trong công thức (2.3) có tỷ lệ thích hợp.

Ngoài ra, nội dung chương còn giới thiệu về bộ dữ liệu đồ án sử dụng cùng với môi trường thử nghiệm và các tiêu chí đánh giá. Trong chương tiếp theo, đồ án sẽ tập trung trình bày về phương pháp xây dựng hệ thống tư vấn khách sạn ứng dụng các kết quả đã đạt được trong chương này.

## CHƯƠNG 4: PHÁT TRIỂN ỨNG DỤNG HỆ TƯ VẤN KHÁCH SẠN

Nội dung trong Chương 4 trình bày mục đích xây dựng hệ thống cùng với Django, công nghệ được đồ án sử dụng để phát triển ứng dụng. Ngoài ra, các biểu đồ ca sử dụng, phân tích và thiết kế cũng được trình bày trong chương này. Nội dung Chương 4 được phân chia như sau:

1. Tổng quan hệ thống
2. Phân tích hệ thống
3. Thiết kế hệ thống
4. Giao diện một số chức năng hệ thống
5. Kết luận

### 4.1. TỔNG QUAN HỆ THỐNG

#### 4.1.1. Giới thiệu hệ thống

Hệ thống đánh giá và tư vấn khách sạn hỗ trợ người dùng đăng các bài đánh giá, tìm kiếm và khám phá thông tin về khách sạn. Ngoài ra, với mục tiêu cá nhân hóa thông tin người dùng, hệ thống cung cấp dịch vụ quản lý tài khoản. Mỗi người dùng với tài khoản của mình có thể thực hiện tìm kiếm, chỉnh sửa thông tin tài khoản và những thông tin này sẽ là cơ sở để hệ thống tư vấn khách sạn một cách chính xác nhất. Ngoài ra, các chức năng giúp người dùng có thể tương tác với nhau thông qua các bài đăng đánh giá như: bình luận, thích cũng được phát triển.

#### 4.1.2. Công nghệ sử dụng

Django là một framework bậc cao của Python có thể thúc đẩy việc phát triển phần mềm một cách nhanh chóng, rõ ràng. Django tập trung lớn những vấn đề phát triển Web. Vì vậy, chỉ cần tập trung xây dựng các phần chức năng của hệ thống mà không cần bắt đầu từ những công việc nhỏ.

Những lợi thế của Django:

1. **Hoàn thiện:** Django bao gồm hàng tá tính năng bổ sung mà bạn có thể sử dụng để xử lý các tác vụ phát triển web thông thường. Django chăm sóc xác thực người dùng, quản trị nội dung, sơ đồ trang web, nguồn cấp dữ liệu RSS và nhiều tác vụ khác - ngay lập tức.
2. **Đa năng:** Django có thể được dùng để xây dựng hầu hết các loại website, từ hệ thống quản lý nội dung, cho đến các trang mạng xã hội hay web tin tức. Nó có thể làm việc với framework phía người dùng, và chuyển nội dung hầu hết các loại format (HTML, RESS, JSON, XML, ...).
3. **Bảo mật:** Django rất coi trọng vấn đề bảo mật và giúp các nhà phát triển tránh được nhiều lỗi bảo mật phổ biến, chẳng hạn như SQL Injection, cross-site script, giả mạo yêu cầu cross-site và clickjacking. Hệ thống xác thực người dùng của Django cung cấp một cách an toàn để quản lý tài khoản và mật khẩu của người dùng.

4. **Mở rộng linh hoạt:** Một số trang web lớn: Instagram, Spotify, ... sử dụng khả năng mở rộng quy mô nhanh chóng và linh hoạt của Django để đáp ứng nhu cầu lưu lượng truy cập lớn nhất.
5. **Cực kỳ linh hoạt:** Django được viết bằng Python, nó có thể chạy đa nền tảng, có nghĩa là nhà phát triển không bị ràng buộc bởi một nền tảng máy chủ cụ thể. Django được hỗ trợ tốt ở nhiều nhà cung cấp hosting, họ sẽ cung cấp hạ tầng và tài liệu cụ thể cho hosting web Django.
6. **Để bảo trì:** Django khuyến khích việc tái sử dụng code. Điều này hỗ trợ hệ thống giảm một lượng mã nguồn đáng kể, hệ thống cũng dễ dàng bảo trì hơn.

## 4.2. PHÂN TÍCH HỆ THỐNG

### 4.2.1. Xây dựng biểu đồ ca sử dụng

#### Xác định và mô tả các tác nhân

- *Người dùng:* người dùng có thể đăng ký tài khoản để truy cập vào hệ thống và sử dụng các chức năng được cho phép.
- *Quản trị viên:* quản trị viên được cung cấp tài khoản cấp cao để truy cập vào hệ thống và có quyền quản lý thông tin tài khoản, khách sạn, bài đánh giá.

#### Xác định và mô tả ca sử dụng

- U1: Đăng nhập: các tác nhân đăng nhập hệ thống
- U2: Đăng xuất: các tác nhân thoát khỏi hệ thống.
- U3: Đăng ký tài khoản: người dùng đăng ký tài khoản để có thể truy cập vào hệ thống
- U4: Đăng bài đánh giá: người dùng đăng bài đánh giá về 1 khách sạn
- U5: Bình luận: người dùng bình luận trong 1 bài viết
- U6: Thích: người dùng thực hiện tương tác với bài viết
- U7: Theo dõi tài khoản: người dùng theo dõi tài khoản khác
- U8: Tìm kiếm khách sạn: người dùng tìm kiếm thông tin về khách sạn
- U9: Tìm kiếm tài khoản: người dùng tìm kiếm thông tin về tài khoản
- U10: Khám phá khách sạn: người dùng nhận được các tư vấn khách sạn có thể phù hợp với bản thân từ hệ thống
- U11: Đổi ảnh đại diện: người dùng đổi ảnh đại diện của tài khoản
- U12: Đổi mật khẩu: người dùng đổi mật khẩu của tài khoản.
- U13: Quản lý tài khoản: quản trị viên thực hiện các chức năng như: tạo mới, cập nhật, vô hiệu hóa, tìm kiếm với dữ liệu tài khoản người dùng
- U14: Quản lý bài đánh giá: quản trị viên thực hiện các chức năng như: tạo mới, cập nhật, vô hiệu hóa, tìm kiếm với dữ liệu bài đánh giá
- U15: Quản lý bài khách sạn: quản trị viên thực hiện các chức năng như: tạo mới, cập nhật, vô hiệu hóa, tìm kiếm với dữ liệu khách sạn



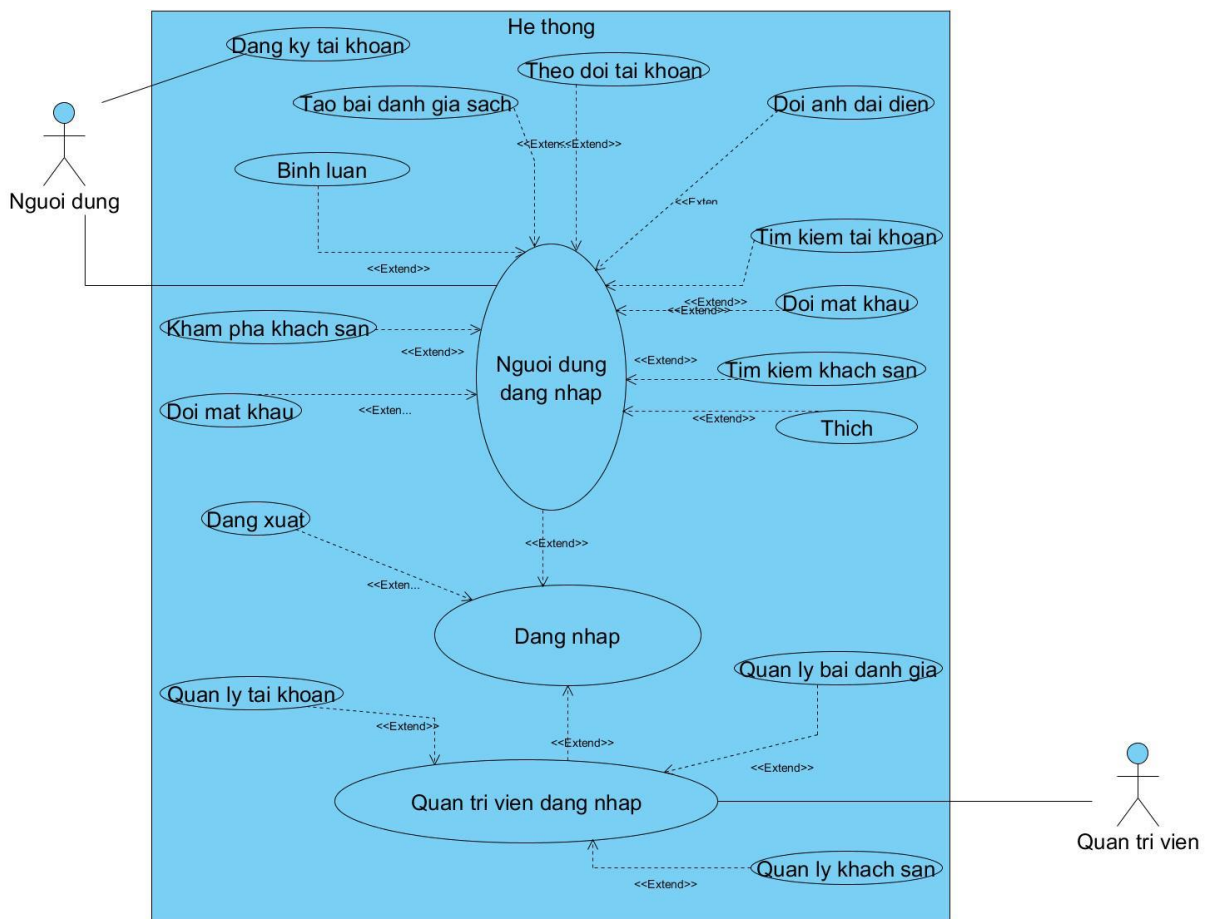
- U16: Huấn luyện mô hình tư vấn: quản trị viên thực hiện huấn luyện mô hình tư vấn khách sạn với dữ liệu mới.

### Khảo sát các ca sử dụng

Khi truy cập vào hệ thống, yêu cầu người dùng phải đăng nhập vào hệ thống (U1) trước khi có thể sử dụng các chức năng của hệ thống. Để có tài khoản truy cập vào hệ thống, người dùng cần phải đăng ký tài khoản (U3).

Sau khi đăng nhập vào hệ thống, người dùng có thể tùy chọn sử dụng các chức năng của hệ thống: đăng bài đánh giá (U4), bình luận (U5), thích các bài viết khác (U6), theo dõi tài khoản khác (U7), tìm kiếm thông tin khách sạn (U8), tìm kiếm thông tin tài khoản (U9), khám phá khách sạn (U10), đổi ảnh đại diện (U11), đổi mật khẩu tài khoản của mình (U12). Đối với quản trị viên, sau khi đăng nhập vào hệ thống với tài khoản được cấp, họ có thể tùy chọn sử dụng các chức năng: quản lý tài khoản (U13), quản lý bài đánh giá (U13), quản lý khách sạn (U15), huấn luyện mô hình tư vấn (U16).

### Biểu đồ ca sử dụng



**4.2.2. Kịch bản ca sử dụng****Đăng bài đánh giá**

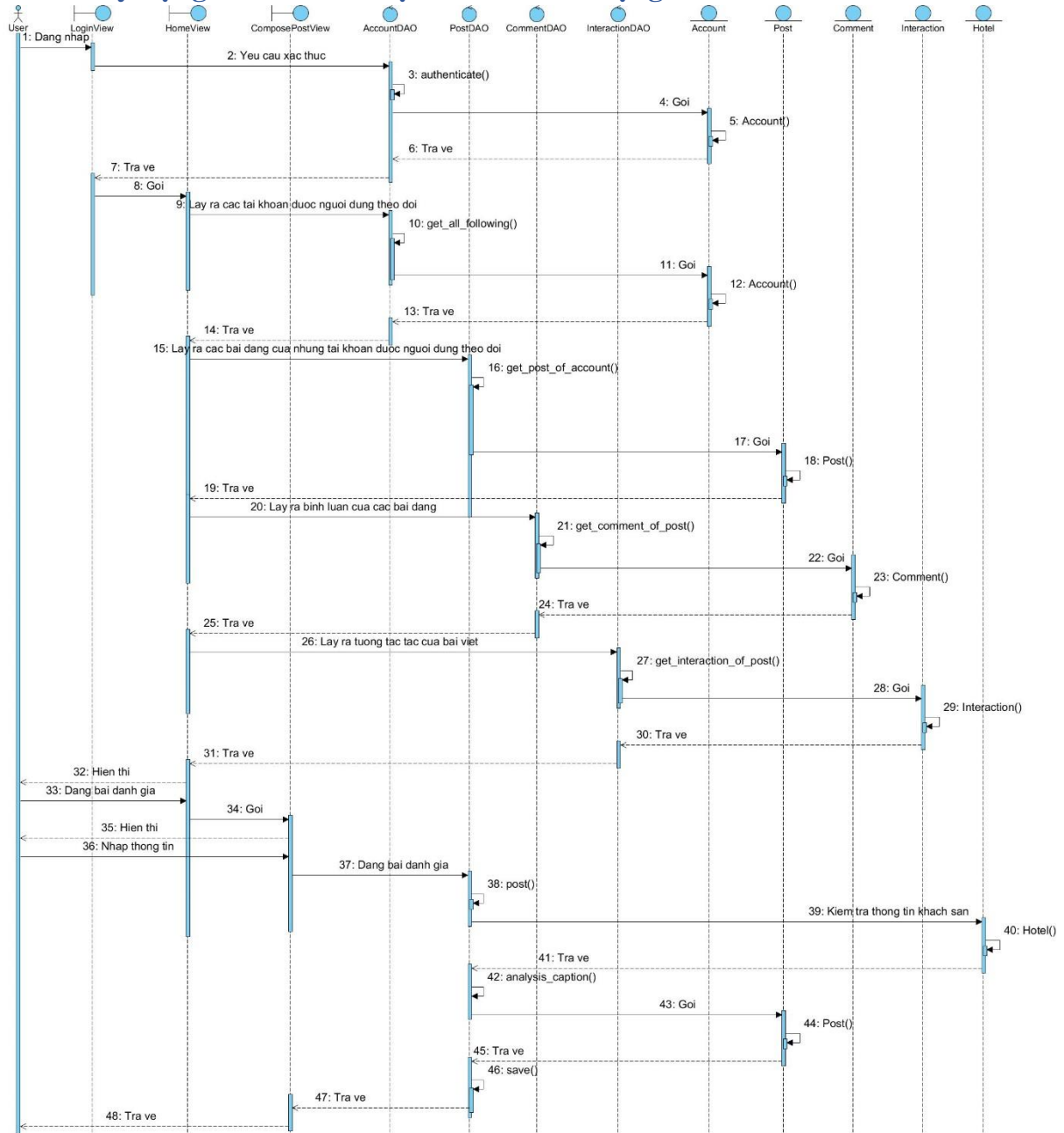
<b>Tên Use Case</b>	Đăng bài đánh giá
<b>Tác nhân</b>	Người dùng hệ thống
<b>Tiền điều kiện</b>	Người dùng đã thực hiện đăng nhập
<b>Hậu điều kiện</b>	Người dùng đăng bài đánh giá thành công
<b>Kịch bản chuẩn:</b> <ol style="list-style-type: none"> <li>1. Người dùng bấm đăng bài đánh giá tại giao diện trang chủ</li> <li>2. Hệ thống hiển thị giao diện đăng bài đánh giá ComposeView cho người dùng</li> <li>3. Người dùng nhập các thông tin cho bài đánh giá và bấm đăng</li> <li>4. Hệ thống hiển thị giao diện trang chủ.</li> </ol>	
<b>Ngoại lệ:</b> <ol style="list-style-type: none"> <li>2.1 Người dùng không nhập điểm đánh giá <ol style="list-style-type: none"> <li>2.1.1 Người dùng không nhập điểm và bấm đăng</li> <li>2.1.2 Hệ thống báo yêu cầu nhập đầy đủ thông tin</li> </ol> </li> <li>2.2 Người dùng không nhập lời bình luận đánh giá <ol style="list-style-type: none"> <li>2.2.1 Người dùng không nhập lời bình luận và bấm đánh giá</li> <li>2.2.2 Hệ thống báo yêu cầu nhập đầy đủ thông tin</li> </ol> </li> <li>2.3 Người dùng không nhập ảnh mô tả <ol style="list-style-type: none"> <li>2.3.1 người dùng không nhập ảnh mô tả</li> <li>2.3.2 Hệ thống báo yêu cầu tải lên ảnh mô tả</li> </ol> </li> </ol>	

**Khám phá khách sạn**

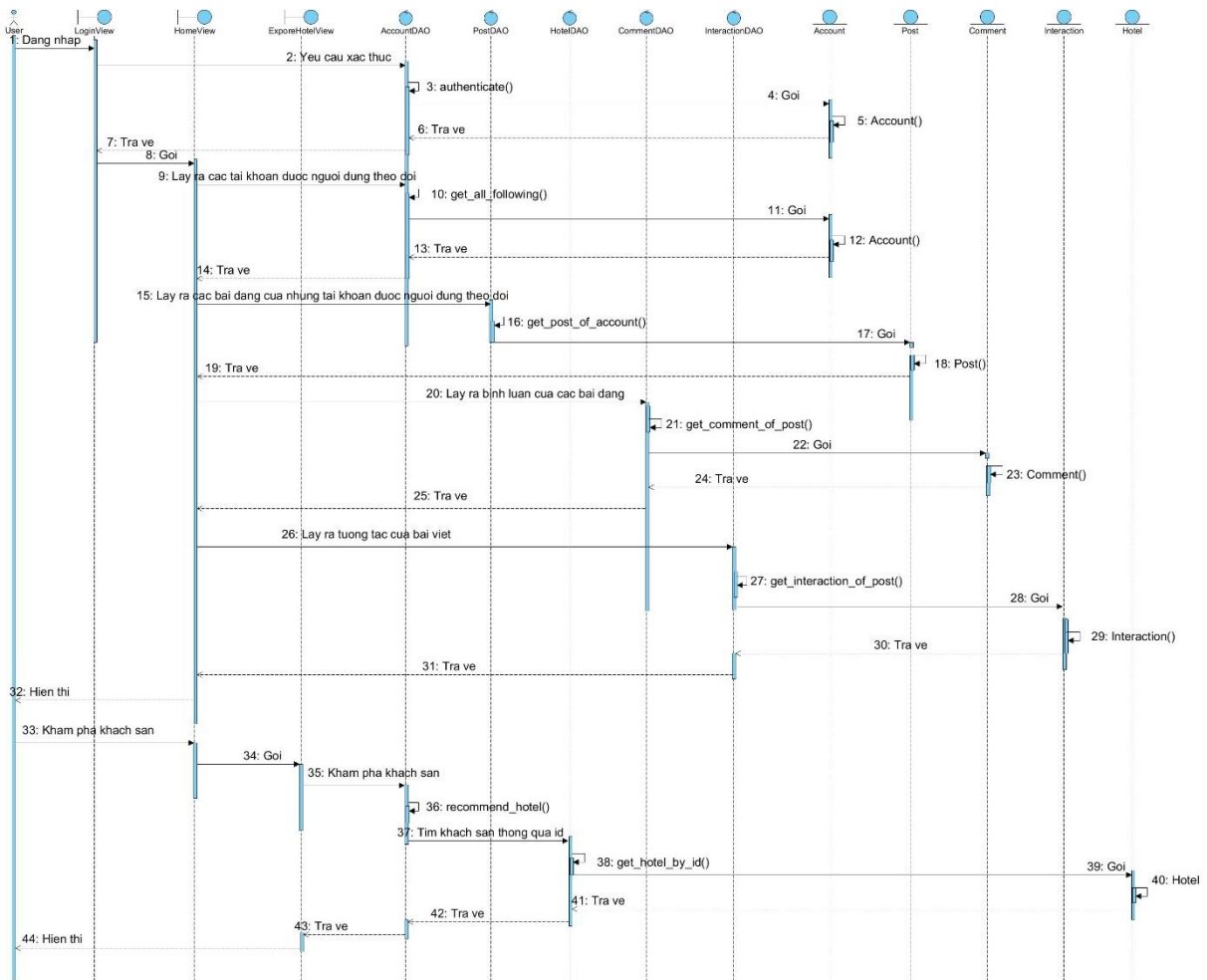
<b>Tên Use Case</b>	Khám phá khách sạn
<b>Tác nhân</b>	Người dùng hệ thống
<b>Tiền điều kiện</b>	Người dùng đã thực hiện đăng nhập
<b>Hậu điều kiện</b>	Người dùng khám phá thành công
<b>Kịch bản chuẩn:</b> <ol style="list-style-type: none"> <li>1. Người dùng A bấm vào nút khám phá khách sạn tại giao diện trang chủ</li> <li>2. Hệ thống trả về giao diện khám phá khách sạn gồm thông tin những khách sạn được tư vấn cho người dùng</li> <li>3. Người dùng bấm vào 1 khách sạn được tư vấn</li> <li>4. Hệ thống trả về giao diện chi tiết khách sạn</li> </ol>	



## 4.2.3. Xây dựng biểu đồ tuần tự của các ca sử dụng



Hình 4.1: Biểu đồ tuần tự ca sử dụng đăng bài đánh giá khách sạn



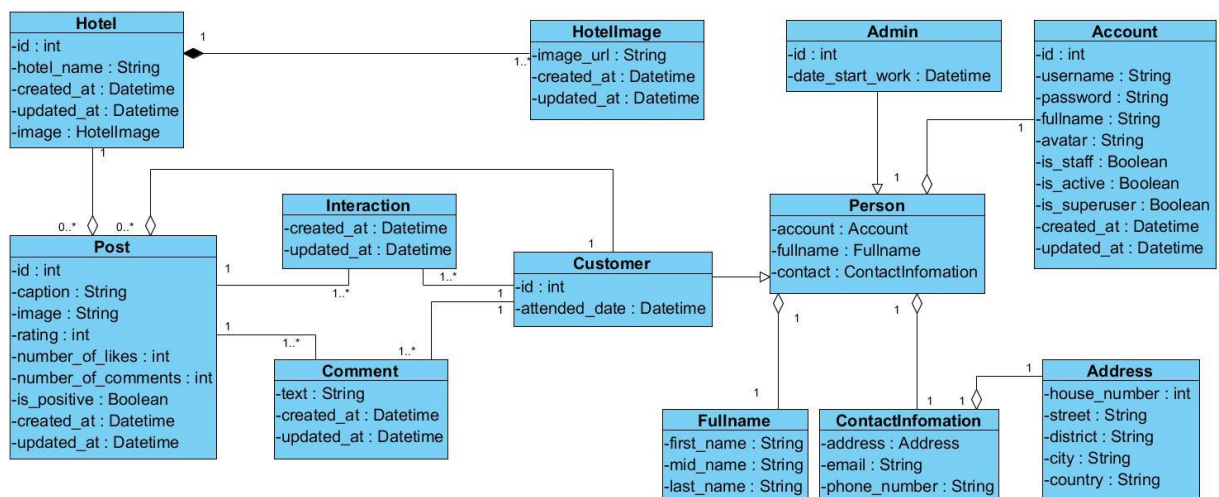
Hình 4.2: Biểu đồ tuần tự ca sử dụng khám phá khách sạn

#### 4.2.4. Xây dựng biểu đồ lớp phân tích

Hotel	Định nghĩa	Chứa các thuộc tính cơ bản của đối tượng là Khách sạn
	Thuộc tính	<ul style="list-style-type: none"> <li>id: mã khách sạn</li> <li>hotel_name: tên khách sạn</li> <li>created_at: thời gian thêm vào hệ thống</li> <li>updated_at: thời gian cập nhật thông tin</li> </ul>
HotelImage	Định nghĩa	Chứa các thuộc tính cơ bản của đối tượng Ảnh khách sạn. Việc tách thành lớp HotelImage phục vụ cho việc quản lý dễ dàng hơn.
	Thuộc tính	<ul style="list-style-type: none"> <li>image_url: đường dẫn ảnh</li> <li>created_at: thời gian thêm vào hệ thống</li> <li>updated_at: thời gian cập nhật thông tin</li> </ul>
Post	Định nghĩa	Chứa các thuộc tính cơ bản của đối tượng Post
	Thuộc tính	<ul style="list-style-type: none"> <li>id: mã bài đăng</li> <li>caption: mô tả</li> </ul>

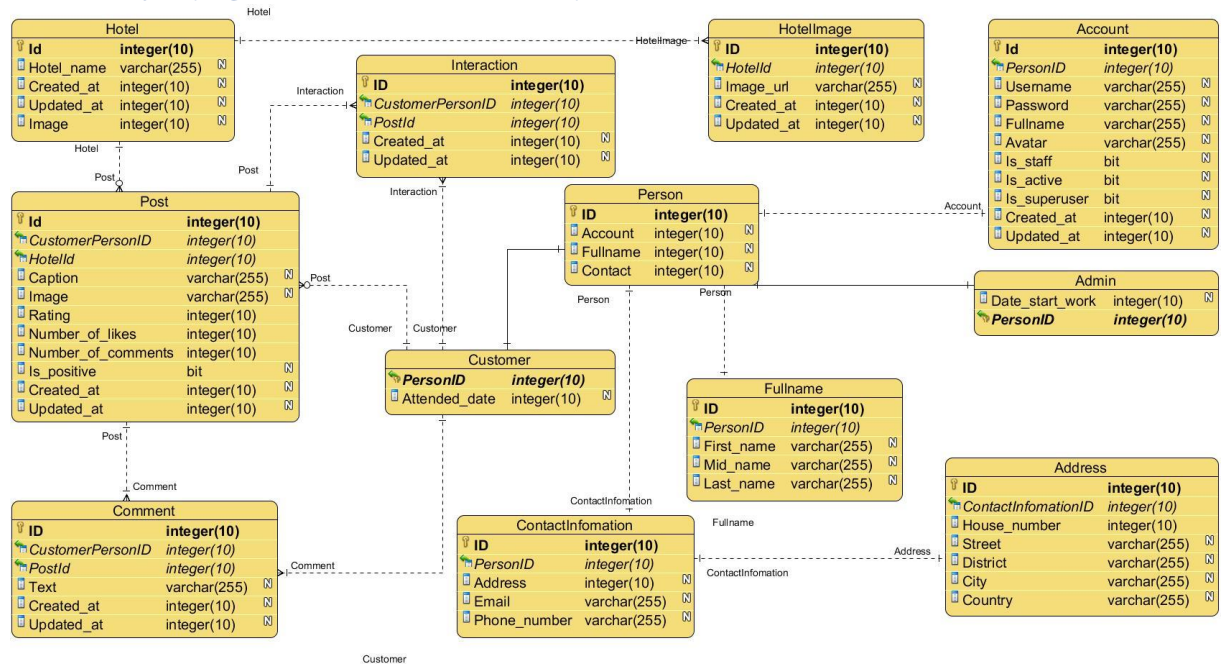
		<ul style="list-style-type: none"> <li>• image: ảnh mô tả</li> <li>• rating: điểm đánh giá</li> <li>• number_of_likes: số lượng người thích</li> <li>• number_of_comments: số lượng bình luận</li> <li>• is_postive: tích cực hay tiêu cực</li> <li>• created_at: thời gian thêm vào hệ thống</li> <li>• updated_at: thời gian cập nhật thông tin</li> </ul>
Account	Định nghĩa	Chứa các thuộc tính cơ bản của đối tượng Account
	Thuộc tính	<ul style="list-style-type: none"> <li>• id: mã tài khoản</li> <li>• username: tên đăng nhập</li> <li>• password: mật khẩu</li> <li>• avatar: ảnh đại diện</li> <li>• is_staff: nhân viên hay không</li> <li>• is_active: tài khoản còn hoạt động không</li> <li>• is_superuser: có phải siêu tài khoản không</li> <li>• created_at: thời gian thêm vào hệ thống</li> <li>• updated_at: thời gian cập nhật thông tin</li> </ul>
Interaction	Định nghĩa	Là lớp liên kết giữa Account và Post, dùng để lưu trữ thông tin khi người dùng thực hiện thích bài viết.
	Thuộc tính	<ul style="list-style-type: none"> <li>• created_at: thời gian thêm vào hệ thống</li> <li>• updated_at: thời gian cập nhật thông tin</li> </ul>
Comment	Định nghĩa	Là lớp liên kết giữa Account và Post, dùng để lưu trữ thông tin khi người dùng thực hiện bình luận bài viết
	Thuộc tính	<ul style="list-style-type: none"> <li>• text: nội dung bình luận</li> <li>• created_at: thời gian thêm vào hệ thống</li> <li>• updated_at: thời gian cập nhật thông tin</li> </ul>
Address	Định nghĩa	Là lớp chứa các thuộc tính cơ bản của địa chỉ nhà
		<ul style="list-style-type: none"> <li>• house_number: số nhà</li> <li>• street: phố</li> <li>• district: quận, huyện</li> <li>• city: thành phố</li> <li>• country: đất nước</li> </ul>
Fullname	Định nghĩa	Là lớp chứa các thuộc tính cơ bản của tên
		<ul style="list-style-type: none"> <li>• first_name: tên gọi</li> <li>• mid_name: tên đệm</li> <li>• last_name: họ</li> </ul>
ContactInformation	Định nghĩa	Là lớp chứa các thông tin liên lạc

		<ul style="list-style-type: none"> <li>• address: địa chỉ nhà</li> <li>• email: địa chỉ email</li> <li>• phone_number: số điện thoại</li> </ul>
Person	Định nghĩa	Là lớp chứa các thuộc tính cơ bản của người
		<ul style="list-style-type: none"> <li>• account: tài khoản</li> <li>• fullname: tên đầy đủ</li> <li>• contact: thông tin liên lạc</li> </ul>
Admin	Định nghĩa	Là lớp chứa thông tin của quản trị viên
		<ul style="list-style-type: none"> <li>• date_start_work: ngày bắt đầu làm việc</li> </ul>
Customer	Định nghĩa	Là lớp chứa thông tin cơ bản của khách hàng
		<ul style="list-style-type: none"> <li>• attended_date: ngày tham gia hệ thống</li> </ul>



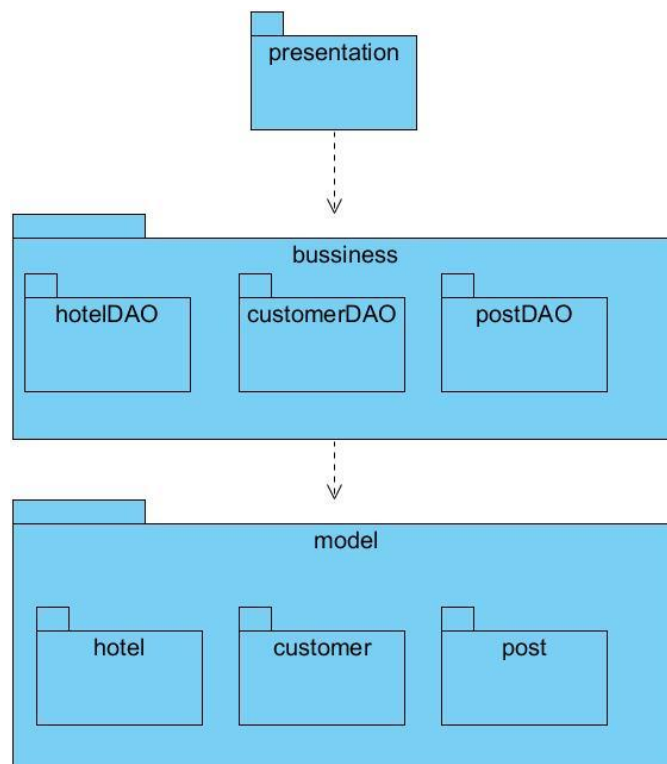
Hình 4.3: Biểu đồ lớp phân tích

## 4.2.5. Xây dựng biểu đồ mô hình dữ liệu



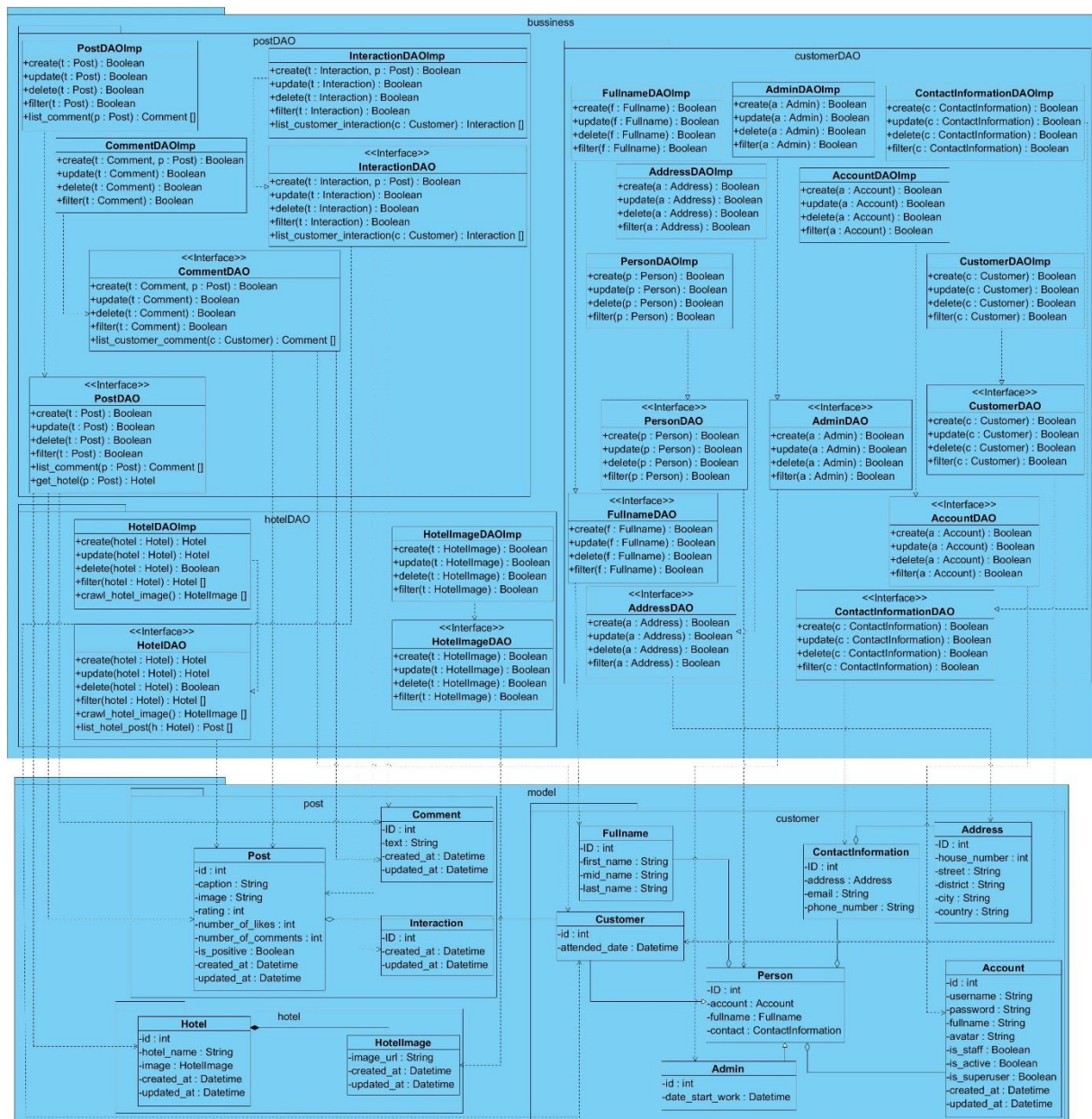
Hình 4.4: Biểu đồ mô hình dữ liệu

## 4.3. THIẾT KẾ HỆ THỐNG



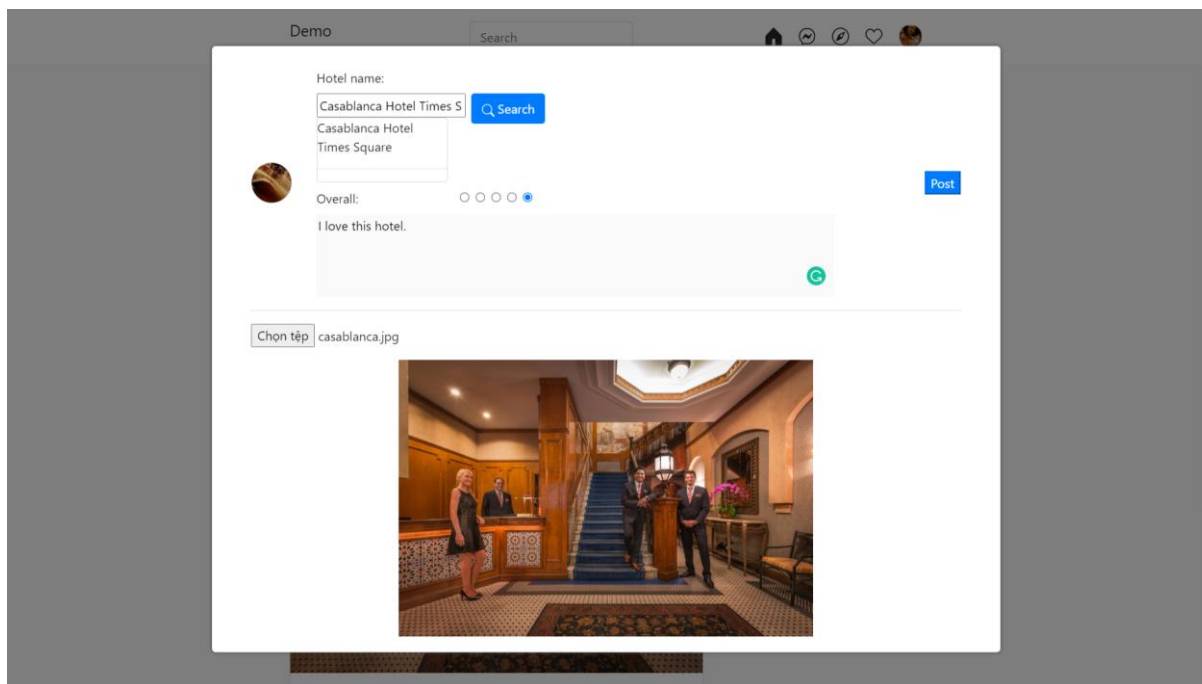
Hình 4.5: Biểu đồ gói của hệ thống



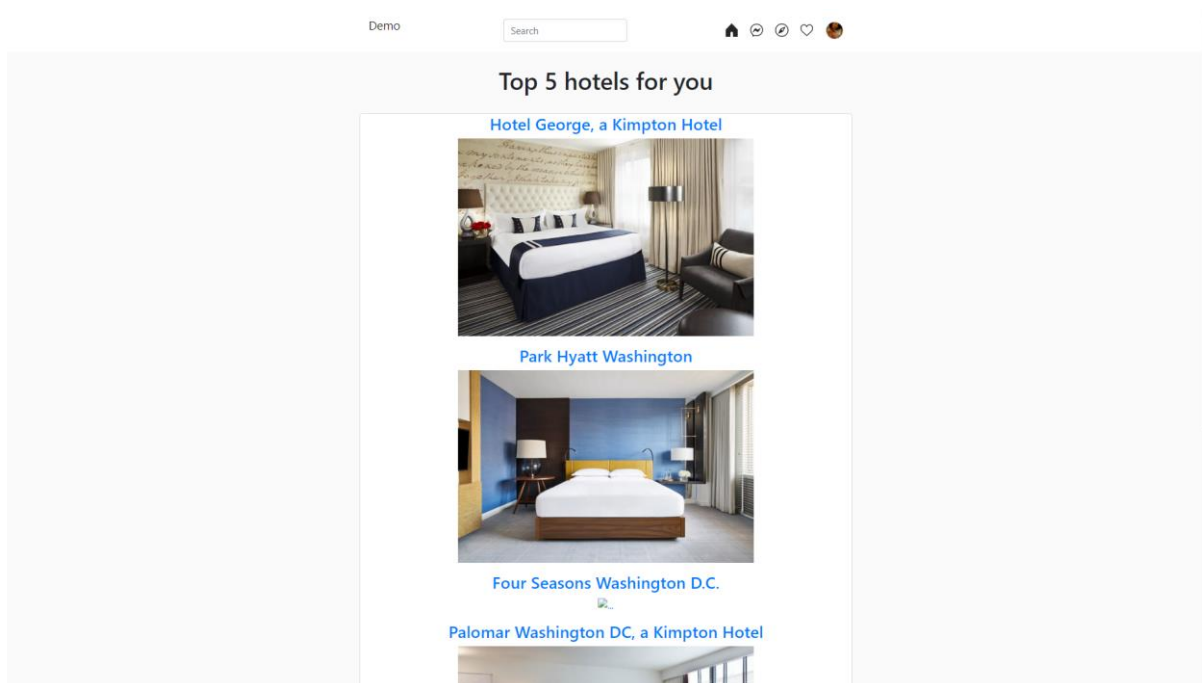


Hình 4.6: Biểu đồ lớp thiết kế

#### 4.4. GIAO DIỆN MỘT SỐ CHỨC NĂNG HỆ THỐNG



Hình 4.7: Giao diện đăng bài đánh giá khách sạn



Hình 4.8: Giao diện tư vấn khách sạn

#### 4.5. KẾT LUẬN

Trong chương này, đồ án đã trình bày tổng quan về hệ thống và công nghệ được sử dụng. Các bước xây dựng, phân tích thiết kế hệ thống cũng được đồ án trình bày trong chương này.

## KẾT LUẬN

### 1. Những kết quả đã đạt được

Với mục tiêu đã đề ra, đồ án đã đi sâu nghiên cứu và khảo sát các vấn đề liên quan đến bài toán tư vấn khách sạn trên các trang thương mại điện tử dựa trên hành vi đánh giá và hành vi bình luận. Dưới đây là những kết quả đồ án đã đạt được:

- Trình bày và áp dụng một số thuật toán học có giám sát và kỹ thuật biểu diễn dữ liệu văn bản cho bài toán Phân loại quan điểm người dùng.
- Phát biểu và xây dựng mô hình xử lý bài toán tư vấn khách sạn dựa trên mô hình dữ liệu kết hợp từ dữ liệu bình luận và dữ liệu đánh giá.
- Áp dụng các kỹ thuật và kết quả nghiên cứu để phát triển hệ thống đánh giá và tư vấn khách sạn.

### 2. Định hướng phát triển trong tương lai

Trong quá trình thực hiện đồ án, không tránh khỏi một số hạn chế do điều kiện về mặt thời gian và trình độ của sinh viên. Vì vậy, trong tương lai, sinh viên dự định thực hiện tiếp các công việc sau:

- Ngoài việc chỉ sử dụng dữ liệu bình luận và dữ liệu đánh giá, đồ án có thể sử dụng thêm dữ liệu như: hành vi thích, hành vi chia sẻ, hành vi tìm kiếm, và các kỹ thuật liên quan đến Trust [13] để thực hiện tư vấn.
- Dự đoán các chủ đề trong bình luận của người dùng để tư vấn khách sạn chính xác hơn. Ví dụ, người dùng thường xuyên nhắc tới tiêu chí sạch sẽ trong các bình luận thì sẽ tư vấn khách sạn có điểm đánh giá tiêu chí này cao cho người dùng.
- Thực hiện nghiên cứu và khảo sát kỹ thuật tư vấn dựa trên mạng nơ-ron [14].

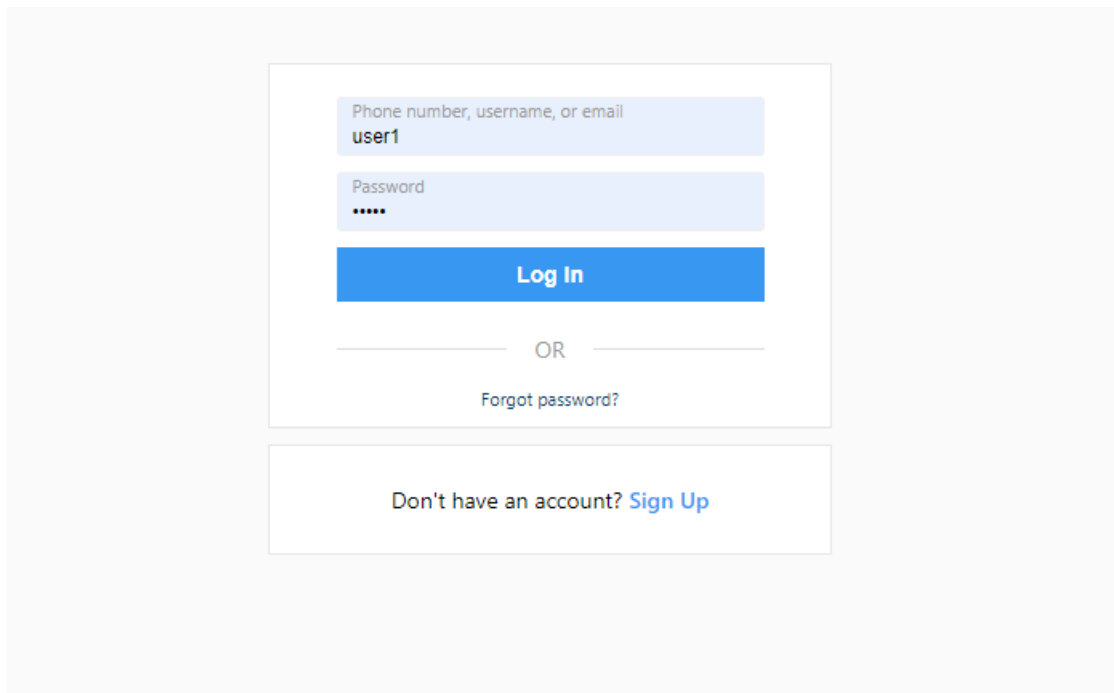


**DANH MỤC TÀI LIỆU THAM KHẢO**

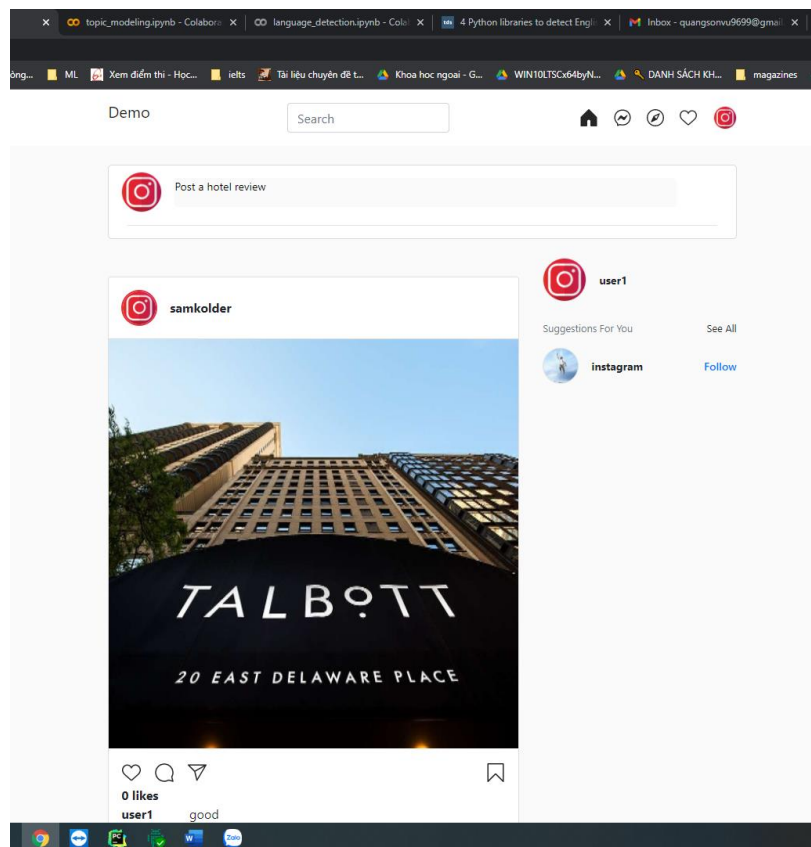
- [1] Edmunds, Angela and Morris, Anne, "The problem of information overload in business organisations: a review of the literature," *International journal of information management*, vol. 20, no. 1, pp. 17--28, 2000.
- [2] Davidson, James and Liebold, Benjamin and Liu, Junning and Nandy, Palash and Van Vleet, Taylor and Gargi, Ullas and Gupta, Sujoy and He, Yu and Lambert, Mike and Livingston, Blake and others, "The YouTube video recommendation system," *Proceedings of the fourth ACM conference on Recommender systems*, pp. 293--296, 2010.
- [3] Isinkaye, Folasade Olubusola and Folajimi, Yetunde O and Ojokoh, Bolande Adefowoke, "Recommendation systems: Principles, methods and evaluation," *Egyptian informatics journal*, vol. 16, no. 3, pp. 261--273, 2015.
- [4] Mohamed, Marwa Hussien and Khafagy, Mohamed Helmy and Ibrahim, Mohamed Hasan, "Recommender systems challenges and solutions survey," *2019 International Conference on Innovative Trends in Computer Engineering (ITCE)*, pp. 149--155, 2019.
- [5] Goyani, Mahesh and Chaurasiya, Neha, "A Review of Movie Recommendation System," *ELCVIA: electronic letters on computer vision and image analysis*, vol. 19, no. 3, pp. 18--37, 2020.
- [6] V. H. Tiệp, *Machine Learning cơ bản*, 2020.
- [7] Yang, Luming and Xu, Min and Xing, Lin, "Exploring the core factors of online purchase decisions by building an E-Commerce network evolution model," *Journal of Retailing and Consumer Services*, vol. 64, 2022.
- [8] Karimi, Mozghan and Jannach, Dietmar and Jugovac, Michael, "News recommender systems - Survey and roads ahead," *Information Processing & Management*, vol. 54, no. 6, pp. 1203--1227, 2018.
- [9] Nguyễn Hà Nam, Nguyễn Trí Thành, Hà Quang Thụy, *Giáo trình khai phá dữ liệu*, Nhà xuất bản Hà Nội, 2016.
- [10] Platt, John, "Sequential minimal optimization: A fast algorithm for training support vector machines," 1998.
- [11] Trần Nguyễn Minh Thư, Phạm Xuân Hiền, "Các phương pháp đánh giá hệ thống gợi ý," *Tạp chí Khoa học Trường Đại học Cần Thơ*, vol. 42, pp. 18--27, 2016.

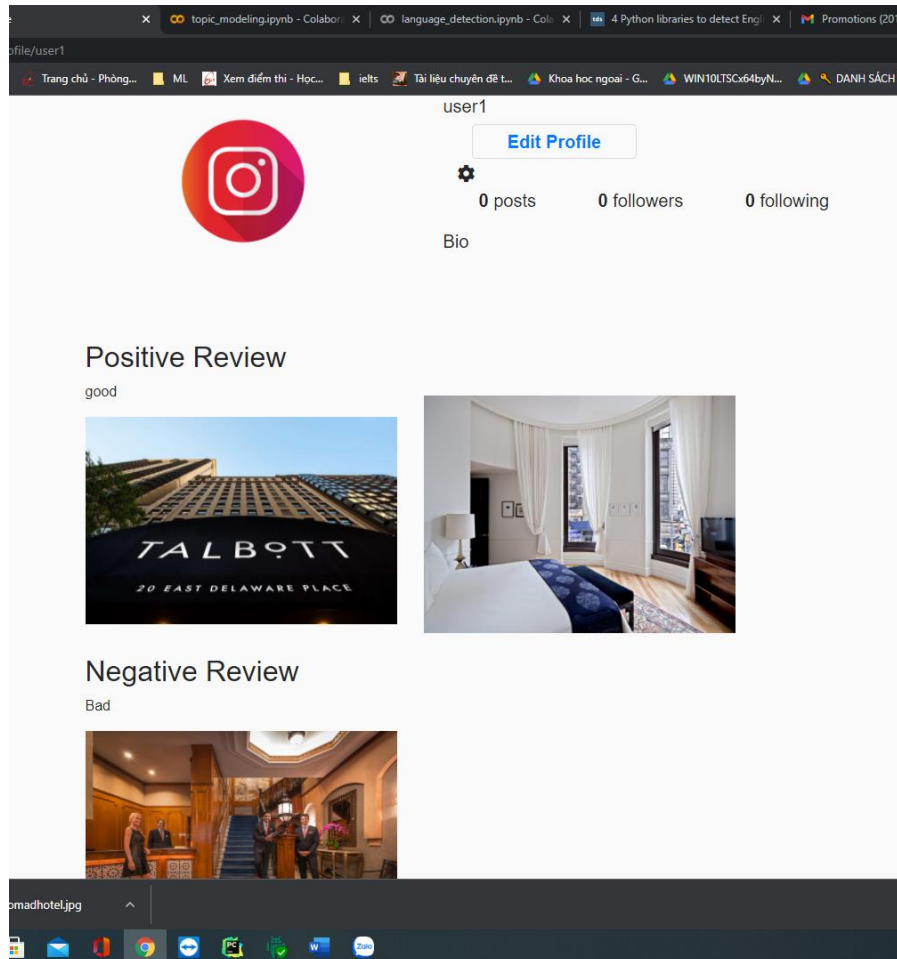
- [12] Salam Patrous, Ziad and Najafi, Safir, "Evaluating Prediction Accuracy for Collaborative Filtering Algorithms in Recommender Systems," 2016.
- [13] Tran, Dinh Que and Pham, Phuong Thanh, "Integrating interaction and similarity threshold of user's interests for topic trust computation," *Southeast Asian Journal of Sciences*, vol. 7, no. 1, pp. 28--35, 2019.
- [14] He, Xiangnan, et al., "Neural collaborative filtering," *Proceedings of the 26th international conference on world wide web*, pp. 173--182, 2017.
- [15] Valdiviezo-Diaz, Priscila and Ortega, Fernando and Cobos, Eduardo and Lara-Cabrera, Raúl, "A collaborative filtering approach based on Naive Bayes classifier," *IEEE Access*, vol. 7, pp. 108581--108592, 2019.
- [16] Koren, Yehuda and Bell, Robert and Volinsky, Chris, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, pp. 30--37, 2009.
- [17] Schafer, J Ben and Frankowski, Dan and Herlocker, Jon and Sen, Shilad, "Collaborative filtering recommender systems," in *The adaptive web*, Springer, 2007, pp. 291--324.
- [18] Ma, Hao and Zhou, Tom Chao and Lyu, Michael R and King, Irwin, "Improving recommender systems by incorporating social contextual information," *ACM Transactions on Information Systems (TOIS)*, vol. 29, no. 2, pp. 1--23, 2011.
- [19] H{\a}ubl, Gerald and Trifts, Valerie, "Consumer decision making in online shopping environments: The effects of interactive decision aids," *Marketing science*, vol. 19, no. 1, pp. 4--21, 2000.
- [20] Laudon, Kenneth C and Traver, Carol Guercio, *E-commerce*, Pearson Boston, MA, 2013.

## PHỤ LỤC

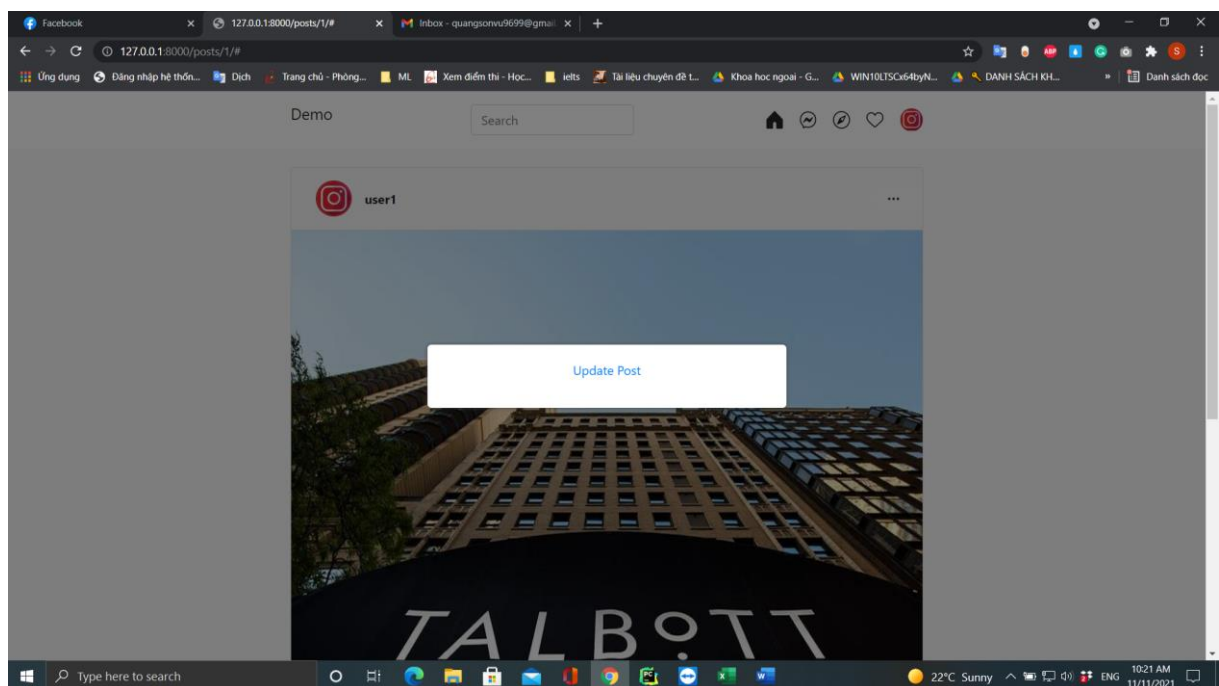


A login form interface with a light gray background. It features a white rectangular box containing two input fields: the first is labeled 'Phone number, username, or email' and contains the text 'user1'; the second is labeled 'Password' and contains six dots. Below these fields is a blue 'Log In' button. Underneath the button is a horizontal line with the word 'OR' in the center. Below the line is a link that says 'Forgot password?'. At the bottom of the white box is a link that says 'Don't have an account? Sign Up'.

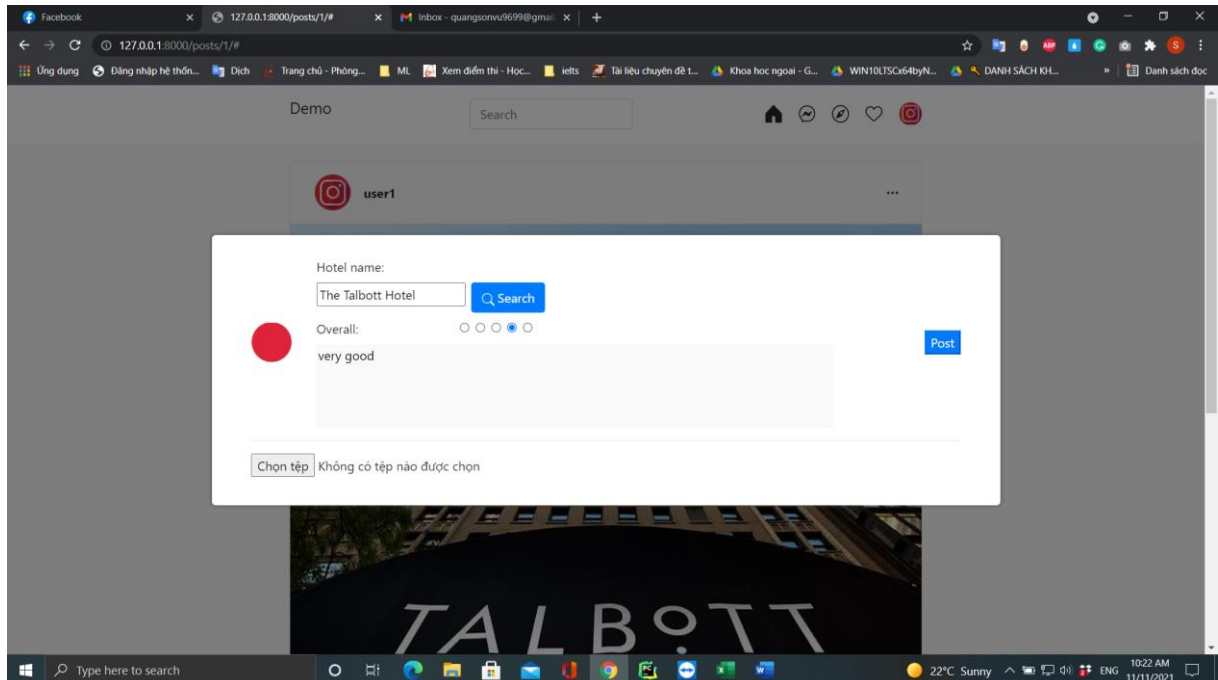
*Hình PL. 1: Giao diện đăng nhập**Hình PL. 2: Giao diện trang chủ*



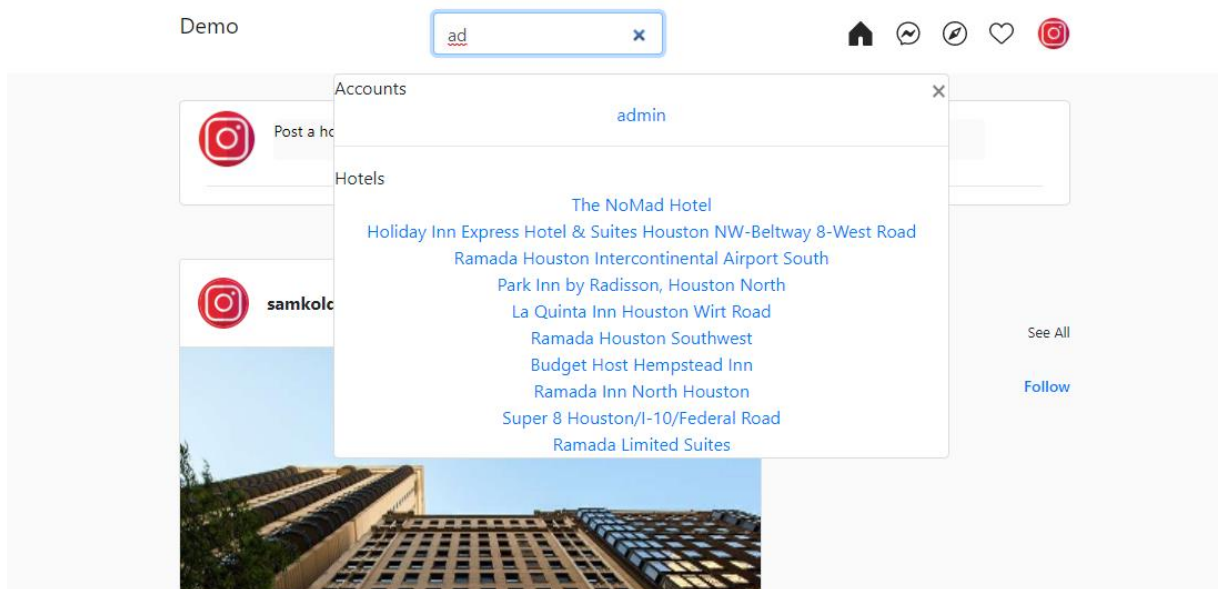
Hình PL. 3: Giao diện trang cá nhân



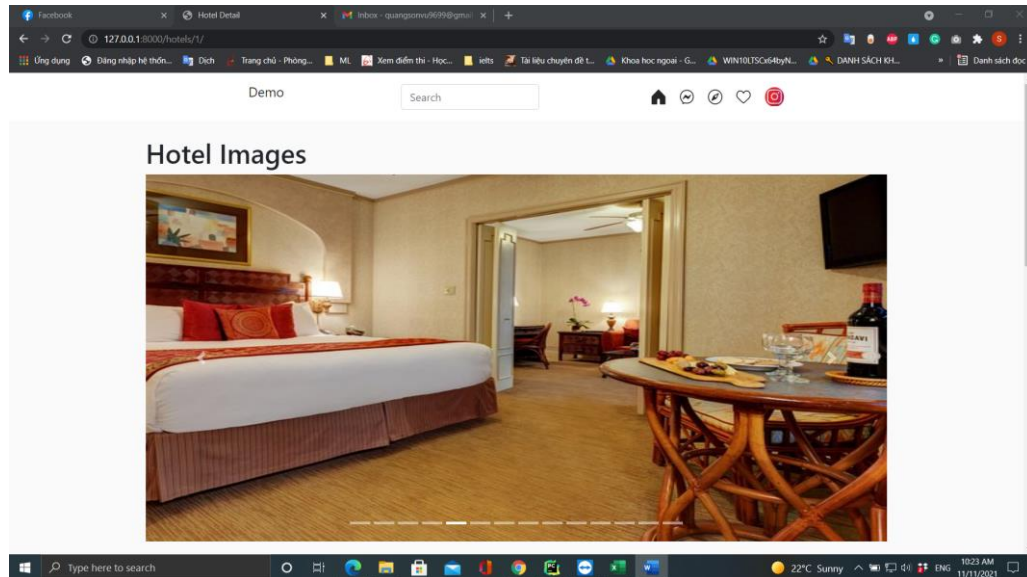
Hình PL. 4: Giao diện chỉnh sửa bài viết 1



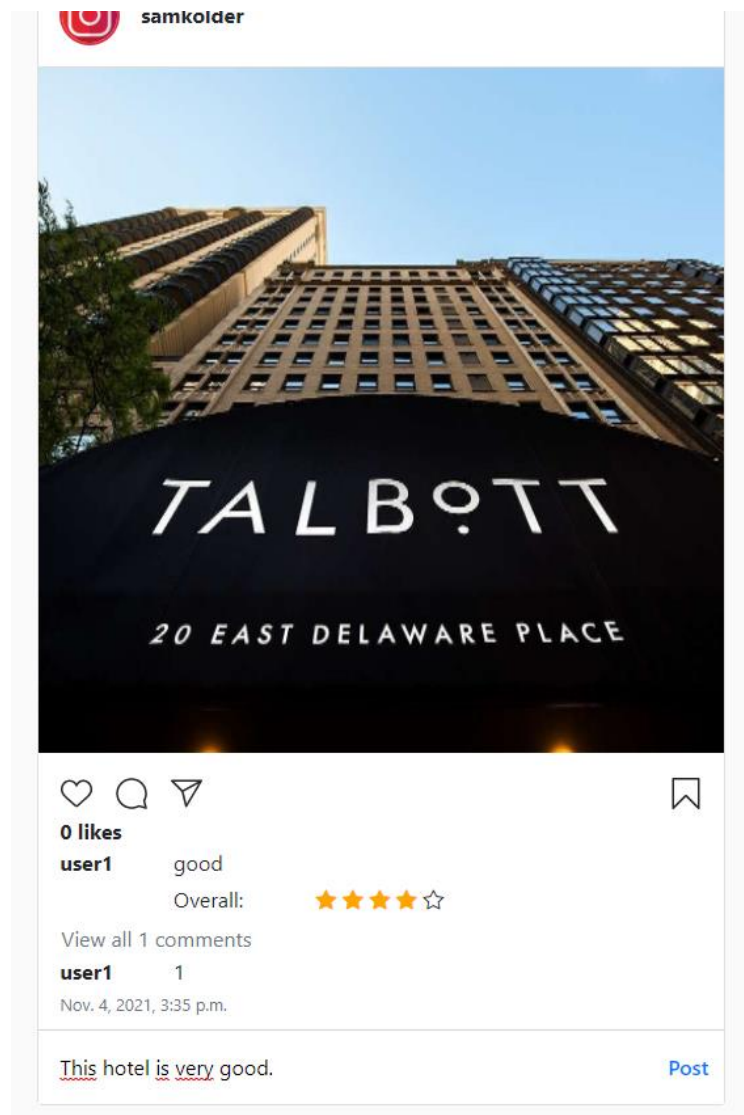
Hình PL. 5: Giao diện chỉnh sửa bài viết 2



Hình PL. 6: Giao diện kết quả tìm kiếm



Hình PL. 7: Giao diện chi tiết khách sạn



Hình PL. 8: Giao diện thích và bình luận