

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**  
**TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT**  
**KHOA HỆ THÔNG TIN THÔNG TIN**

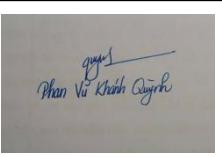
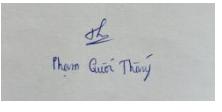
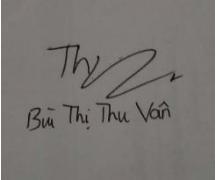


**ĐỒ ÁN CUỐI KỲ**  
**MÔN: KỸ THUẬT LẬP TRÌNH**  
**ĐỀ TÀI**  
**PHÁT TRIỂN ỨNG DỤNG PHÂN TÍCH CẢM XÚC ĐÁNH GIÁ**  
**KHÁCH SẠN DỰA TRÊN NLP VÀ DEEP LEARNING**

**GVHD: Th.S Nguyễn Quang Phúc**  
**Mã HP: 242BIE501901**

*Thành phố Hồ Chí Minh, ngày 25 tháng 3 năm 2025*

### THÀNH VIÊN NHÓM 3

STT	Họ và tên	MSSV	Đóng góp	Chữ ký
1	Nguyễn Lê Minh Tài	K224101336	100%	
2	Lê Hữu Đăng	K234060688	100%	
3	Phan Vũ Khánh Quỳnh	K234060723	100%	
4	Phạm Quốc Thắng	K234060727	100%	
5	Bùi Thị Thu Vân	K234060739	100%	

## LỜI CẢM ƠN

Lời đầu tiên, nhóm chúng em xin gửi lời cảm ơn chân thành đến thầy - ThS. Nguyễn Quang Phúc - giảng viên môn Kỹ thuật lập trình. Trong quá trình tìm hiểu và học tập, nhóm chúng em đã nhận được sự giảng dạy, hướng dẫn rất tận tình và tâm huyết từ thầy cũng như sự giúp đỡ, hỗ trợ về mặt kiến thức. Nhờ vậy nhóm chúng em đã có thể tích lũy thêm nhiều điều mới và có cái nhìn sâu sắc hơn về môn học này. Từ những kiến thức cũng như những gợi ý được truyền đạt từ thầy, nhóm chúng em xin trình bày lại những gì mình đã tích lũy và tìm hiểu được trong suốt quá trình vừa qua.

Tuy nhiên do giới hạn kiến thức và khả năng lý luận của nhóm còn nhiều hạn chế, kính mong sự chỉ dẫn và đóng góp từ Thầy để đồ án của nhóm chúng em được hoàn thiện hơn.

Kính chúc thầy thật nhiều sức khỏe và đạt được nhiều thành công trong cuộc sống. Nhóm em xin chân thành cảm ơn!

*Nhóm 3 KTLT*

## MỤC LỤC

THÀNH VIÊN NHÓM 3 .....	1
LỜI CẢM ƠN .....	2
DANH MỤC HÌNH ẢNH .....	6
DANH MỤC BẢNG BIỂU .....	8
DANH MỤC TÊN VIẾT TẮT .....	9
TÓM TẮT .....	12
CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI .....	13
1.1 Động lực .....	13
1.2 Mục tiêu nghiên cứu .....	14
1.3 Đối tượng và phạm vi nghiên cứu .....	16
1.3.1 Đối tượng .....	16
1.3.2 Phạm vi .....	16
1.4 Các công cụ sử dụng .....	16
1.5 Cấu trúc dự án .....	16
CHƯƠNG 2: CƠ SỞ LÝ THUYẾT .....	18
2.1 Xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP) .....	18
2.1.1. Phân loại .....	18
2.1.2. Tầm quan trọng của NLP .....	18
2.1.3. Công nghệ trong NLP .....	18
2.1.4. Cấu trúc của NLP .....	19
2.1.5. Ứng dụng thực tiễn .....	19
2.2 Các mô hình học máy .....	20
2.2.1 Bernoulli Naive Bayes (BernoulliNB) .....	20
2.2.2 Random Forest .....	21
2.2.3 Support Vector Classifier (SVC) .....	22

2.2.4 Logistic Regression .....	23
2.3 Mô hình học sâu LSTM .....	25
CHƯƠNG 3: PHƯƠNG PHÁP ĐỀ XUẤT .....	29
3.1 Thu thập dữ liệu .....	29
3.1.1 Tập dữ liệu .....	29
3.1.2 Tiền xử lý dữ liệu .....	30
3.1.3 Trực quan hóa dữ liệu .....	35
3.2 Quy trình xây dựng mô hình NLP .....	38
3.3. Kiến trúc ứng dụng và luồng UI/UX .....	41
CHƯƠNG 4: MÔ HÌNH VÀ QUY TRÌNH THỰC HIỆN .....	43
4.1 Thiết kế cơ sở dữ liệu .....	43
4.1.1 Entities and Categories .....	43
4.1.2 Mô tả quy tắc nghiệp vụ .....	46
4.1.3. Ràng buộc lực lượng của mô tả mối quan hệ .....	46
4.1.4. Thiết kế cơ sở dữ liệu vật lý .....	47
4.2 Sơ đồ Use Case .....	49
4.2.1 Sơ đồ Use Case .....	49
4.2.2 Mô tả Use Case .....	51
4.3 Thiết kế giao diện người dùng .....	53
4.3.1 Giao diện đăng nhập .....	53
4.3.2 Giao diện sử dụng .....	56
CHƯƠNG 5: KẾT QUẢ THỰC NGHIỆM .....	59
5.1 Đánh giá mô hình Sentiment Analysis .....	59
5.2 Trải nghiệm người dùng .....	62
5.2.1 Màn hình đăng nhập .....	62

5.2.2 Quên mật khẩu .....	63
5.2.3 Đăng ký tài khoản .....	64
5.2.4 Trang chủ .....	65
CHƯƠNG 6: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN TRONG TƯƠNG LAI .....	79
6.1 Kết luận .....	79
6.2 Hạn chế và hướng phát triển trong tương lai .....	79
TAI LIỆU THAM KHẢO .....	81

## DANH MỤC HÌNH ẢNH

Hình 2.1: Trợ lý AI .....	20
Hình 2.2: Thuật toán Random Forest trong học máy .....	22
Hình 2.3: Hàm Sigmoid .....	24
Hình 2.4: Mô tả cấu trúc mạng LSTM và cấu trúc của một mô-đun của nó .....	26
Hình 3.1: Các bước tiền xử lý dữ liệu .....	30
Hình 3.2: Minh họa việc loại bỏ stopwords .....	32
Hình 3.3: Mô tả Stemming .....	33
Hình 3.4: Tổng quan phân bố lượt đánh giá .....	35
Hình 3.5: Tổng quan độ dài theo mức độ đánh giá .....	35
Hình 3.6: Tỷ lệ các mức đánh giá khách sạn .....	36
Hình 3.7: Tỷ lệ các loại hình du lịch .....	37
Hình 3.8: Xếp hạng số lượng bài đánh giá của các khách sạn .....	38
Hình 3.9: Quy trình xây dựng mô hình NLP .....	39
Hình 3.10: Quy trình phát triển ứng dụng .....	41
Hình 4.1: Database ERD .....	43
Hình 4.2: Use case diagram .....	50
Hình 4.3: Giao diện đăng nhập .....	53
Hình 4.4: Giao diện quên mật khẩu .....	54
Hình 4.5: Giao diện đăng ký tài khoản .....	55
Hình 4.6: Giao diện Visualize Reviews .....	56
Hình 4.7: Giao diện View Data Customers .....	57
Hình 4.8: Giao diện View Data Hotels .....	57
Hình 4.9: Giao diện View Data Customers .....	57
Hình 4.10: Giao diện View Data Comments .....	57
Hình 5.1: Màn hình đăng nhập .....	62
Hình 5.2: Màn hình quên mật khẩu .....	63
Hình 5.3: Màn hình đăng ký tài khoản .....	64
Hình 5.4: Trang chủ .....	65
Hình 5.5: Khách sạn Orleans Hotel and Casino .....	67
Hình 5.6: Minh họa về thống kê đơn giản .....	67

Hình 5.7: Minh họa về phân bố điểm đánh giá .....	68
Hình 5.8: Minh họa về tỷ lệ đánh giá tích cực và tiêu cực .....	69
Hình 5.9: Minh họa về xu hướng điểm đánh giá .....	70
Hình 5.10: Minh họa về phân bố khách hàng theo quốc gia .....	71
Hình 5.11: Minh họa về phân bố theo loại phòng .....	72
Hình 5.12: Minh họa về phân phối thời gian lưu trú .....	73
Hình 5.13: Hiển thị dữ liệu khách hàng .....	74
Hình 5.14: Hiển thị dữ liệu khách sạn .....	75
Hình 5.15: Hiển thị danh sách quốc gia .....	76
Hình 5.16: Hiển thị danh sách các bình luận đánh giá của khách hàng về các khách sạn .....	77

## DANH MỤC BẢNG BIỂU

Bảng 4.1: Các thực thể của Database .....	43
Bảng 4.2: Các thuộc tính của thực thể Userapp .....	44
Bảng 4.3: Các thuộc tính của thực thể Customer .....	44
Bảng 4.4: Các thuộc tính của thực thể Hotel .....	45
Bảng 4.5: Các thuộc tính của thực thể Comment .....	45
Bảng 4.6: Các thuộc tính của thực thể Country .....	45
Bảng 4.7: Mối quan hệ giữa các thực thể .....	47
Bảng 4.8: Cơ sở vật lý thực thể Country .....	47
Bảng 4.9: Cơ sở vật lý thực thể Customer .....	47
Bảng 4.10: Cơ sở vật lý thực thể Hotel .....	48
Bảng 4.11: Cơ sở vật lý thực thể Comment .....	49
Bảng 4.12: Mô tả giao diện đăng nhập .....	53
Bảng 4.13: Mô tả giao diện quên mật khẩu .....	54
Bảng 4.14: Mô tả giao diện đăng ký .....	55
Bảng 4.15: Mô tả giao diện Visualize Reviews .....	57
Bảng 4.16: Mô tả giao diện View Data .....	58
Bảng 5.1: So sánh các thuật toán .....	59
Bảng 5.2: So sánh Logistic Regression và LSTM .....	60

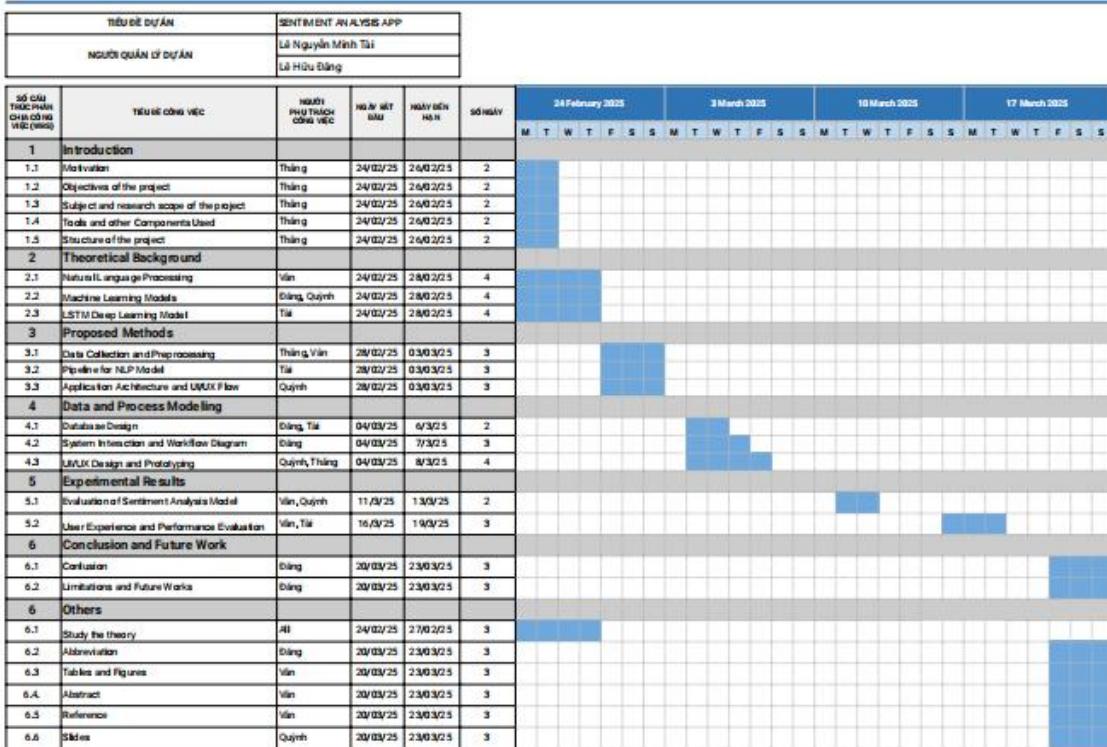
## DANH MỤC TÊN VIỆT TẮT

Từ viết tắt	Ý nghĩa
ADASYN	Adaptive Synthetic Sampling (Lấy mẫu tổng hợp thích ứng)
AI	Artificial Intelligence (Trí tuệ nhân tạo)
CNN	Convolutional Neural Network (Mạng nơ-ron tích chập)
EDA	Exploratory Data Analysis (Phân tích dữ liệu khám phá)
ERD	Entity-Relationship Diagram (Sơ đồ thực thể - quan hệ)
LIME	Local Interpretable Model-agnostic Explanations (Giải thích mô hình cục bộ độc lập)
LSTM	Long Short-Term Memory (Bộ nhớ ngắn dài hạn)
ML	Machine Learning (Học máy)
MT	Machine Translation (Dịch máy)
NER	Named Entity Recognition (Nhận diện thực thể có tên)
NLG	Natural Language Generation (Sinh ngôn ngữ tự nhiên)
NLP	Natural Language Processing (Xử lý ngôn ngữ tự nhiên)
POS	Part of Speech (Phân loại từ loại)
RNN	Recurrent Neural Network (Mạng nơ-ron hồi quy)
SHAP	SHapley Additive exPlanations (Giải thích cộng tính Shapley)
SMOTE	Synthetic Minority Over-sampling Technique (Kỹ thuật tổng hợp dữ

	liệu thiểu số)
SVC	Support Vector Classifier (Bộ phân loại vector hỗ trợ)
SVM	Support Vector Machine (Máy vector hỗ trợ)
TF-IDF	Term Frequency - Inverse Document Frequency (Tần suất thuật ngữ - Tần suất nghịch đảo tài liệu)
XAI	Explainable Artificial Intelligence (Trí tuệ nhân tạo giải thích được)

## GANNT CHART

BIỂU ĐỒ GANTT



## TÓM TẮT

Với sự phát triển mạnh mẽ của các nền tảng trực tuyến và mạng xã hội, lượng phản hồi từ khách hàng ngày càng gia tăng, ảnh hưởng đáng kể đến ngành du lịch và khách sạn. Do đó, nhu cầu về một ứng dụng phân tích cảm xúc tự động để hỗ trợ doanh nghiệp trong việc đánh giá ý kiến khách hàng ngày càng trở nên quan trọng. Dự án này tập trung vào việc phát triển một ứng dụng phân tích cảm xúc dựa trên Xử lý ngôn ngữ tự nhiên (NLP) và Học máy (ML) nhằm khai thác dữ liệu đánh giá khách sạn trên nền tảng Agoda. Ứng dụng được thiết kế để thu thập, tiền xử lý và phân tích dữ liệu, từ đó xác định mức độ hài lòng của khách hàng và đưa ra những thông tin hữu ích cho doanh nghiệp. Trong nghiên cứu này, các mô hình học máy truyền thống như Naïve Bayes, Random Forest, Logistic Regression, Support Vector Classifier (SVC) cùng với mô hình học sâu LSTM đã được thử nghiệm. Các mô hình được đánh giá dựa trên các tiêu chí độ chính xác, độ nhạy, độ đặc hiệu và điểm F1. Kết quả thực nghiệm cho thấy, mô hình LSTM kết hợp với Word Embedding đạt hiệu suất tốt nhất trong việc phân tích cảm xúc từ các đánh giá của khách hàng. Ứng dụng được tích hợp với cơ sở dữ liệu và cung cấp giao diện thân thiện, hỗ trợ trực quan hóa dữ liệu để doanh nghiệp dễ dàng khai thác thông tin. Nghiên cứu này không chỉ giúp tối ưu hóa quá trình phân tích cảm xúc mà còn tạo tiền đề cho việc mở rộng ứng dụng trong nhiều lĩnh vực khác như thương mại điện tử và chăm sóc khách hàng.

## CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI

### 1.1 Động lực

Sự phát triển mạnh mẽ của các nền tảng trực tuyến và mạng xã hội đã làm gia tăng đáng kể mức độ tương tác của khách hàng đối với các dịch vụ khách sạn. Những phản hồi này không chỉ tạo ra một lượng dữ liệu khổng lồ ở cả dạng có cấu trúc và phi cấu trúc mà còn đóng vai trò quan trọng trong việc hình thành nhận thức và quyết định của khách hàng tiềm năng.<sup>1</sup> Ngành du lịch và khách sạn đặc biệt chịu ảnh hưởng lớn từ những phản hồi này. Theo báo cáo của Google, hơn 80% khách du lịch nghiên cứu trực tuyến trước khi đặt phòng và dành nhiều thời gian để tìm hiểu đánh giá từ những người đi trước.<sup>2</sup> Do đó, việc phân tích cảm xúc từ các phản hồi trực tuyến không chỉ giúp doanh nghiệp hiểu rõ hơn về tâm lý khách hàng mà còn hỗ trợ cải thiện chất lượng dịch vụ, tăng cường lòng trung thành và tạo ra lợi thế cạnh tranh bền vững trên thị trường.

Sự cần thiết của công nghệ NLP còn được minh chứng qua việc giúp doanh nghiệp nhanh chóng phân tích hàng loạt các đánh giá, từ đó nhận diện được các điểm mạnh, điểm yếu của sản phẩm cũng như xu hướng tiêu dùng hiện hành. Các nghiên cứu cho thấy rằng, nhờ vào việc áp dụng các thuật toán học máy và học sâu, các ứng dụng phân tích cảm xúc đã đạt được hiệu suất cao, giúp giảm thiểu sai sót và tối ưu hóa quá trình xử lý dữ liệu. Những mô hình như CNN, LSTM hay các mô hình lai kết hợp giữa CNN và RNN đã được áp dụng thành công, đem lại độ chính xác trong việc nhận diện các cảm xúc trong các tình huống phức tạp.<sup>3</sup>

Hơn nữa, trong một thị trường cạnh tranh ngày càng khốc liệt, việc nắm bắt kịp thời những phản hồi từ khách hàng đóng vai trò quan trọng trong việc điều chỉnh chiến lược kinh doanh. Sự kết hợp giữa phân tích cảm xúc và các phương pháp học máy hiện đại không chỉ giúp dự đoán xu hướng trải nghiệm dịch vụ mà còn tạo ra

<sup>1</sup> Pabel, A. and Prideaux, B. (2016), “Social media use in pre-trip planning by tourists visiting a small regional leisure destination”, Journal of Vacation Marketing, Vol. 22 No. 4, pp. 335-348.

<sup>2</sup> The Telegraph(2013), “Tripadvisor and the issue of trust”, Visited on 2020-01-28.

<sup>3</sup> Punithavathi Rasappan, Manoharan Premkumar, Garima Sinha, & Kumar Chandrasekaran. (2024, May). Transforming sentiment analysis for e-commerce product reviews: Hybrid deep learning model with an innovative term weighting and feature selection.

những lợi thế cạnh tranh bền vững cho doanh nghiệp. Qua đó, các công cụ phân tích cảm xúc giúp doanh nghiệp không chỉ cải thiện chất lượng dịch vụ mà còn tối ưu hóa quá trình chăm sóc khách hàng, từ đó xây dựng niềm tin và tăng cường lòng trung thành của khách hàng đối với thương hiệu.<sup>4</sup>

Từ những lý do trên, động lực phát triển dự án phần mềm phân tích cảm xúc cho nhân viên được hình thành dựa trên nhu cầu thực tế của thị trường thương mại điện tử nói chung và thị trường khách sạn nói riêng hiện nay. Việc sử dụng công nghệ NLP để tự động thu thập, xử lý và phân tích các nhận xét của khách hàng không chỉ giúp tiết kiệm thời gian, chi phí mà còn tạo ra một công cụ hỗ trợ đắc lực trong việc ra quyết định kinh doanh, giúp mở ra nhiều cơ hội phát triển, từ việc cải thiện chất lượng dịch vụ đến xây dựng các chiến lược tiếp thị hiệu quả.

## 1.2 Mục tiêu nghiên cứu

Dự án của nhóm hướng đến việc phát triển một ứng dụng phân tích cảm xúc tự động, nhằm cung cấp cho các doanh nghiệp khách sạn thông tin chi tiết, kịp thời và chính xác về cảm nhận của khách hàng. Mục tiêu chính là giúp doanh nghiệp hiểu rõ mức độ hài lòng, xu hướng đánh giá và những yếu tố ảnh hưởng đến trải nghiệm khách hàng, từ đó điều chỉnh chiến lược dịch vụ phù hợp để nâng cao chất lượng và khả năng cạnh tranh.

Trước hết, ứng dụng được thiết kế để tự động thu thập dữ liệu từ các nền tảng đánh giá khách sạn trực tuyến, sàng lọc và xử lý các phản hồi không hợp lệ. Điều này giúp đảm bảo rằng dữ liệu thu thập được đầy đủ, chính xác và có giá trị thực tiễn cho quá trình phân tích cảm xúc. Các kỹ thuật crawl hiện đại sẽ được áp dụng để thu thập thông tin một cách liên tục và chính xác, giúp doanh nghiệp cập nhật phản hồi của khách hàng theo thời gian thực.

Tiếp theo, một trong những mục tiêu trọng tâm của dự án là phát triển module phân tích cảm xúc dựa trên các thuật toán Xử lý ngôn ngữ tự nhiên (NLP) và Học máy

---

<sup>4</sup> Gaurav Meena, Krishna Kumar Mohbey, & Sunil Kumar. (2023, April). Sentiment analysis on images using convolutional neural networks based Inception-V3 transfer learning approach.

(ML). Module này sẽ xử lý các phản hồi từ khách hàng, phân loại chúng theo các mức độ cảm xúc như tích cực, tiêu cực hoặc trung tính. Thông tin này giúp doanh nghiệp dễ dàng nhận diện các điểm mạnh, điểm yếu trong dịch vụ, từ đó đưa ra các biện pháp cải thiện phù hợp. Những tiến bộ trong NLP đã chứng minh hiệu quả trong việc trích xuất thông tin từ văn bản phi cấu trúc, giúp rút ngắn thời gian phân tích và tăng tính chính xác trong dự đoán cảm xúc.

Ngoài ra, dự án cũng hướng tới việc xây dựng một hệ thống cơ sở dữ liệu tích hợp, giúp lưu trữ, quản lý và truy xuất thông tin phân tích một cách thuận tiện. Hệ thống này sẽ có giao diện thân thiện, hỗ trợ các doanh nghiệp khách sạn trong việc tìm kiếm và xem xét các xu hướng đánh giá từ khách hàng, giúp họ đưa ra quyết định chiến lược nhanh chóng và hiệu quả.

Việc kết hợp giữa các module crawl dữ liệu, phân tích cảm xúc và hệ thống quản lý dữ liệu không chỉ tối ưu hóa quy trình xử lý thông tin mà còn giúp doanh nghiệp nắm bắt xu hướng thị trường một cách toàn diện. Điều này tạo ra lợi thế cạnh tranh cho các khách sạn, giúp họ nâng cao chất lượng dịch vụ, cải thiện trải nghiệm khách hàng và gia tăng tỷ lệ đặt phòng trong môi trường du lịch số hóa hiện nay.

Để hướng tới mục tiêu này, nhóm đặt ra một số câu hỏi nghiên cứu phụ nhằm định hướng cho quá trình phát triển hệ thống:

1. Làm thế nào để áp dụng hiệu quả các thuật toán NLP và ML trong việc phân tích cảm xúc từ bình luận sản phẩm?
2. Làm thế nào để thiết kế module crawl dữ liệu một cách thống nhất để tiền xử lý và phân tích dữ liệu đảm bảo độ chính xác và liên tục khi thu thập các bình luận từ một nền tảng khách sạn?
3. Làm thế nào để tích hợp hệ thống cơ sở dữ liệu với các module phân tích và thu thập dữ liệu, từ đó hỗ trợ nhân viên trong việc đưa ra các chiến lược kinh doanh phù hợp?

Những câu hỏi trên phản ánh các khía cạnh cốt lõi của dự án, từ việc phát triển thuật toán phân tích cảm xúc đến quy trình thu thập dữ liệu và quản lý thông tin.

Phương pháp tiếp cận dự án được xây dựng trên cơ sở tích hợp các công nghệ tiên tiến với quy trình làm việc hiệu quả, nhằm tạo ra một hệ thống có khả năng đáp ứng nhanh chóng những thay đổi của thị trường khách sạn. Việc áp dụng một chiến lược đa chiều không chỉ giúp giải quyết được các vấn đề hiện có mà còn tạo nền tảng cho những cải tiến trong tương lai.

### 1.3 Đối tượng và phạm vi nghiên cứu

#### 1.3.1 Đối tượng

Đối tượng nghiên cứu của dự án là các bình luận đánh giá trên website <https://www.agoda.com/>

#### 1.3.2 Phạm vi

**Phạm vi thời gian:** Hệ thống sẽ thu thập toàn bộ các bình luận sản phẩm trên nền tảng mà không giới hạn khoảng thời gian, từ dữ liệu cũ cho đến mới nhất.

**Phạm vi không gian:** Nghiên cứu bao quát tất cả các bình luận đánh giá sau khi trải nghiệm trên website agoda.

### 1.4 Các công cụ sử dụng

*Qt Design:* Công cụ hỗ trợ thiết kế giao diện người dùng cho các ứng dụng phát triển bằng Qt, giúp xây dựng giao diện trực quan và tối ưu trải nghiệm người dùng.

*Figma:* Nền tảng thiết kế giao diện và nguyên mẫu (UI/UX) trực tuyến, cho phép nhóm làm việc cộng tác, thiết kế và chỉnh sửa giao diện một cách linh hoạt và hiệu quả.

*PyCharm:* Môi trường phát triển tích hợp (IDE) dành cho Python, được nhóm sử dụng để viết, kiểm thử và triển khai mã nguồn. Ngoài ra, PyCharm còn hỗ trợ thu thập dữ liệu từ sàn Agoda, giúp trích xuất và tổng hợp bình luận của khách hàng một cách nhanh chóng và chính xác.

*Draw.io:* Công cụ hỗ trợ thiết kế luồng hoạt động của ứng dụng, trực quan hóa quy trình vận hành nhằm giúp nhóm có cái nhìn rõ ràng hơn trong quá trình phát triển và hoàn thiện đề tài.

### 1.5 Cấu trúc dự án

Dự án này bao gồm sáu chương chính, trình bày chi tiết các khía cạnh khác nhau của dự án, cụ thể như sau:

## ***Chương 1: Tổng quan đề tài***

Chương này giới thiệu tổng quan về dự án, bao gồm động lực nghiên cứu, mục tiêu, đối tượng, phạm vi nghiên cứu và các công cụ được sử dụng.

## ***Chương 2: Cơ sở lý thuyết***

Chương này tập trung vào các lý thuyết nền tảng liên quan đến phân tích cảm xúc, xử lý ngôn ngữ tự nhiên (NLP), học máy (ML) và tổng quan các nghiên cứu trước đây có liên quan, làm cơ sở cho phương pháp tiếp cận của dự án.

## ***Chương 3: Phương pháp đề xuất***

Chương này mô tả quy trình thu thập, lưu trữ và xử lý dữ liệu, xây dựng mô hình dữ liệu cũng như cách hệ thống tổ chức và phân loại thông tin để tối ưu hóa quá trình phân tích cảm xúc.

## ***Chương 4: Mô hình và quy trình thực hiện***

Chương này trình bày các phương pháp và thuật toán được đề xuất để thu thập, phân tích và xử lý dữ liệu bình luận sản phẩm trên ..., bao gồm các mô hình học máy được áp dụng.

## ***Chương 5: Kết quả thực nghiệm***

Chương này trình bày kết quả thực nghiệm của hệ thống, đánh giá hiệu suất của các mô hình phân tích cảm xúc và so sánh độ chính xác giữa các phương pháp tiếp cận khác nhau.

## ***Chapter 6: Kết luận và hướng phát triển trong tương lai***

Chương này tổng kết những kết quả đạt được của dự án, đánh giá hiệu quả của hệ thống, đồng thời đề xuất các hướng phát triển trong tương lai để cải thiện mô hình phân tích cảm xúc và mở rộng phạm vi ứng dụng.

## CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

### 2.1 Xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP)

Xử lý ngôn ngữ tự nhiên (NLP) là một lĩnh vực thuộc trí tuệ nhân tạo (AI), tập trung vào khả năng giúp máy tính hiểu, diễn giải và tạo ra ngôn ngữ con người một cách tự động. NLP kết hợp giữa khoa học máy tính, ngôn ngữ học và thống kê để xử lý các dữ liệu văn bản và giọng nói. Công nghệ này ngày càng trở nên quan trọng khi lượng dữ liệu văn bản số hóa tăng mạnh, tạo ra nhu cầu xử lý và khai thác thông tin một cách hiệu quả.<sup>5</sup>

#### 2.1.1. Phân loại

NLP được chia thành hai lĩnh vực chính: ngôn ngữ học và khoa học máy tính. Ngôn ngữ học tập trung vào việc hiểu cấu trúc và ý nghĩa của ngôn ngữ, bao gồm các khía cạnh như ngữ âm, âm vị, cú pháp, ngữ nghĩa và ngữ dụng học. Trong khi đó, khoa học máy tính quan tâm đến việc chuyển đổi những kiến thức ngôn ngữ này thành các chương trình máy tính thông qua việc áp dụng các mô hình như hệ thống dựa trên quy tắc, học máy cổ điển và học sâu.<sup>6</sup>

#### 2.1.2. Tầm quan trọng của NLP

Trong thời đại số hóa, khối lượng dữ liệu văn bản và giọng nói ngày càng tăng từ các nguồn như email, mạng xã hội và cuộc gọi dịch vụ khách hàng. NLP giúp tự động hóa việc xử lý và phân tích những dữ liệu này, cho phép doanh nghiệp hiểu rõ hơn về phản hồi của khách hàng, cải thiện dịch vụ và tối ưu hóa hoạt động kinh doanh. Ví dụ, các chatbot sử dụng NLP có thể trả lời tự động các câu hỏi thường gặp, giảm tải cho nhân viên và nâng cao trải nghiệm khách hàng.

#### 2.1.3. Công nghệ trong NLP

Các công nghệ trong NLP rất đa dạng và bao gồm nhiều phương pháp xử lý dữ liệu ngôn ngữ. Một số công nghệ quan trọng trong NLP có thể kể đến như xử lý văn bản (Text Processing), giúp chuẩn hóa dữ liệu thông qua loại bỏ ký tự đặc biệt, chuyển đổi chữ hoa/chữ thường, sửa lỗi chính tả; phân tích cú pháp (Parsing), giúp xác định cấu trúc của câu và mối quan hệ giữa các từ<sup>7</sup>. Nhận dạng thực thể có tên

<sup>5</sup> Jurafsky, D., & Martin, J. H. (2021). *Speech and Language Processing* (3rd ed.). Pearson.

<sup>6</sup> Christopher Manning, Hinrich Schütze, and Prabhakar Raghavan. 2008. Cambridge University Press. Introduction to Information Retrieval.

<sup>7</sup> Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

(Named Entity Recognition - NER) được sử dụng để trích xuất thông tin như tên người, tổ chức, địa điểm từ văn bản.<sup>8</sup> Ngoài ra, các công nghệ như sinh ngôn ngữ tự nhiên (Natural Language Generation - NLG) giúp chuyển đổi dữ liệu thành văn bản có nghĩa, trong khi nhận diện giọng nói (Speech Recognition) cho phép chuyển đổi giọng nói thành văn bản. Một ứng dụng khác là dịch máy (Machine Translation - MT), được sử dụng rộng rãi trong các công cụ như Google Translate.<sup>9</sup>

Một công nghệ quan trọng khác là phân tích cảm xúc (Sentiment Analysis - SA), giúp đánh giá thái độ của người dùng thông qua các văn bản như bình luận, đánh giá sản phẩm và bài đăng trên mạng xã hội. Sentiment Analysis sử dụng các mô hình như Naive Bayes, SVM hoặc mạng nơ-ron sâu để xác định mức độ tích cực, tiêu cực hoặc trung lập của văn bản.<sup>10</sup> Công nghệ này được áp dụng rộng rãi trong marketing, nghiên cứu thị trường và hỗ trợ khách hàng.

#### 2.1.4. Cấu trúc của NLP

Cấu trúc của NLP thường bao gồm bốn giai đoạn chính. Đầu tiên là tiền xử lý dữ liệu, trong đó văn bản được chuẩn hóa thông qua các bước như loại bỏ từ, dừng, tách từ và gán nhãn từ loại. Tiếp theo, hệ thống thực hiện phân tích cú pháp và ngữ nghĩa để hiểu mối quan hệ giữa các từ và ý nghĩa của câu. Sau đó, NLP xử lý ngữ cảnh và suy luận để hiểu sâu hơn về nội dung, đặc biệt quan trọng trong các ứng dụng chatbot hoặc trợ lý ảo. Cuối cùng, hệ thống sinh văn bản hoặc phản hồi dưới dạng ngôn ngữ tự nhiên, giúp tạo ra câu trả lời phù hợp với ngữ cảnh.

#### 2.1.5. Ứng dụng thực tiễn



<sup>8</sup> Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural Architectures for Named Entity Recognition. arXiv preprint arXiv:1603.01360.

<sup>9</sup> Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. ICLR.

<sup>10</sup> Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*

## Hình 2.1: Trợ lý AI<sup>11</sup>

Một trong những ứng dụng phổ biến nhất của NLP là trong các hệ thống chatbot và trợ lý ảo như Siri, Google Assistant và Alexa. Các hệ thống này sử dụng NLP để hiểu và phản hồi các câu hỏi của người dùng một cách tự động, giúp giảm tải công việc cho con người và nâng cao trải nghiệm người dùng.<sup>12</sup> Bên cạnh đó, NLP đóng vai trò quan trọng trong dịch thuật tự động. Các công cụ như Google Translate sử dụng công nghệ dịch máy dựa trên mạng nơ-ron để cải thiện chất lượng bản dịch bằng cách học từ dữ liệu ngũ liệu song ngũ không lồ. Phương pháp này giúp máy dịch hiểu được ngữ cảnh tốt hơn, từ đó tạo ra những bản dịch tự nhiên và chính xác hơn so với các phương pháp dịch máy truyền thống.<sup>13</sup> Còn trong lĩnh vực phân tích cảm xúc, NLP giúp doanh nghiệp đánh giá phản hồi của khách hàng thông qua các bình luận trên mạng xã hội, đánh giá sản phẩm hoặc khảo sát ý kiến. Các mô hình phân tích cảm xúc có thể xác định mức độ tích cực, tiêu cực hoặc trung lập của nội dung, từ đó hỗ trợ các doanh nghiệp điều chỉnh chiến lược tiếp thị và nâng cao chất lượng dịch vụ. Nhờ những tiến bộ trong NLP, các ứng dụng của công nghệ này ngày càng mở rộng, giúp cải thiện đáng kể nhiều khía cạnh của đời sống, từ giao tiếp hàng ngày đến nghiên cứu khoa học và quản lý doanh nghiệp.

Mặc dù NLP đã đạt được nhiều thành tựu đáng kể, công nghệ này vẫn tồn tại một số hạn chế. Một trong những thách thức lớn nhất là độ chính xác chưa cao khi xử lý các ngôn ngữ ít phổ biến do thiếu dữ liệu huấn luyện. Ngoài ra, NLP gặp khó khăn trong việc xử lý các từ hoặc câu có nghĩa đa dạng tùy thuộc vào ngữ cảnh. Hệ thống NLP hiện tại cũng chưa thể suy luận như con người, do đó đôi khi tạo ra câu trả lời không hợp lý hoặc thiếu tính thực tiễn. Những hạn chế này đặt ra nhu cầu cải tiến liên tục trong lĩnh vực NLP để nâng cao khả năng hiểu và xử lý ngôn ngữ tự nhiên.

## 2.2 Các mô hình học máy

### 2.2.1 Bernoulli Naive Bayes (BernoulliNB)

Bernoulli Naive Bayes là một thuật toán phân loại dựa trên Định lý Bayes, giả định rằng các đặc trưng là độc lập với nhau. BernoulliNB triển khai các thuật toán

<sup>11</sup> Kaspersky. (n.d.). *Trợ lý ảo Alexa, Siri và Google Assistant có sử dụng AI không?*. ProGuide. <https://kaspersky.proguide.vn/san-pham-cong-nghe-moi/tro-ly-ao-alex-siri-va-google-assistant-co-su-dung-ai-khong/>

<sup>12</sup> Jurafsky, D., & Martin, J. H. (2021). *Speech and Language Processing* (3rd ed.). Pearson.

<sup>13</sup> Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. ICLR.

huấn luyện và phân loại Naive Bayes cho dữ liệu được phân phối theo phân phối Bernoulli đa biến, tức là có thể có nhiều đặc trưng nhưng mỗi đặc trưng được giả định là một biến nhị phân (Bernoulli, Boolean). Do đó, mô hình này yêu cầu các mẫu dữ liệu được biểu diễn dưới dạng vector đặc trưng nhị phân. Nếu đầu vào không ở dạng nhị phân, một đối tượng BernoulliNB có thể thực hiện chuyển đổi dữ liệu thành dạng nhị phân tùy thuộc vào tham số binarize. Quy tắc quyết định của Bernoulli Naive Bayes dựa trên công thức:

$$P(x_i | y) = P(x_i = 1 | y)x_i + (1 - P(x_i = 1 | y))(1 - x_i)^{14}$$

Trong đó:

$P(x_i | y)$ : là xác suất có điều kiện của  $x_i$  xảy ra với điều kiện  $y$  đã xảy ra i là sự kiện

$x_i$  là giá trị nhị phân 0 hoặc 1

Khác với quy tắc của Naive Bayes phân phối đa thức (MultinomialNB) ở chỗ nó phạt rõ ràng sự vắng mặt của một đặc trưng nếu đặc trưng đó là một chỉ báo quan trọng cho một lớp cụ thể. Trong khi đó, mô hình MultinomialNB sẽ chỉ đơn giản bỏ qua các đặc trưng không xuất hiện.<sup>15</sup>

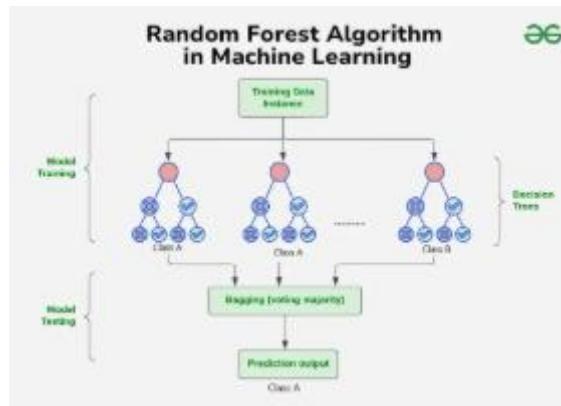
## 2.2.2 Random Forest

Random Forest là một thuật toán học máy thuộc nhóm Ensemble Learning, được phát triển dựa trên mô hình Decision Tree. Thuật toán Random Forest giúp cải thiện độ chính xác của mô hình và giảm hiện tượng quá khớp (overfitting), một vấn đề phổ biến của Decision Tree đơn lẻ. Nhờ vào khả năng tổng hợp kết quả từ nhiều Decision Tree khác nhau, Random Forest có thể xử lý dữ liệu phức tạp, làm việc tốt trên cả các bài toán phân loại (classification) và hồi quy (regression).<sup>16</sup>

<sup>14</sup> GeeksforGeeks. (n.d.). *Bernoulli naive Bayes*. Truy cập từ <https://www.geeksforgeeks.org/bernoulli-naive-bayes/>.

<sup>15</sup> Scikit-learn. (n.d.). *Naive Bayes classifiers*. Scikit-learn. Truy cập từ [https://scikit-learn.org/stable/modules/naive\\_bayes.html](https://scikit-learn.org/stable/modules/naive_bayes.html)

<sup>16</sup> Lê, H. (n.d.). *Rừng ngẫu nhiên (Random Forest)*. Machine Learning Cơ Bản. Truy cập từ [https://machinelearningcoban.com/tabml\\_book/ch\\_model/random\\_forest.html](https://machinelearningcoban.com/tabml_book/ch_model/random_forest.html)



Hình 2.2: Thuật toán Random Forest trong học máy<sup>17</sup>

Như được biểu diễn trên hình ảnh mô họa quy trình huấn luyện của thuật toán Random Forest trong học máy có 2 giai đoạn chính:

**Giai đoạn Huấn Luyện:** Mô hình áp dụng phương pháp bootstrap sampling để tạo ra nhiều tập dữ liệu con từ tập dữ liệu gốc. Mỗi tập con được sử dụng để huấn luyện một Decision Tree riêng biệt. Trong quá trình xây dựng cây, tại mỗi nút phân tách, chỉ một tập hợp con các đặc trưng được chọn ngẫu nhiên để tìm ra tiêu chí phân chia tối ưu, đảm bảo sự đa dạng giữa các cây.<sup>18</sup>

**Giai đoạn Dự Đoán:** Khi có dữ liệu mới, tất cả các Decision Tree đưa ra dự đoán độc lập. Kết quả cuối cùng được tổng hợp thông qua bỏ phiếu đa số (majority voting) trong bài toán phân loại hoặc trung bình cộng (averaging) trong bài toán hồi quy.

### 2.2.3 Support Vector Classifier (SVC)

SVC là một mô hình học máy thuộc họ Support Vector Machines (SVM), được sử dụng chủ yếu cho các bài toán phân loại (classification). SVC hoạt động bằng cách tìm một siêu phẳng (hyperplane) tối ưu trong không gian đặc trưng (feature space) để phân tách các lớp dữ liệu. Mục tiêu của SVC là tối đa hóa khoảng cách lè (margin) giữa các lớp, từ đó giúp mô hình có khả năng tổng quát hóa tốt hơn trên dữ liệu mới.

Cách SVC hoạt động:

<sup>17</sup> GeeksforGeeks. (n.d.). *Random Forest* [Image]. Truy cập từ <https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/>

<sup>18</sup> GeeksforGeeks. (n.d.). *Random forest algorithm in machine learning*. Truy cập từ <https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/>

- Tìm siêu phẳng tối ưu: Mục tiêu của SVC là tìm một siêu phẳng (trong không gian 2D là một đường thẳng) phân chia các điểm dữ liệu thuộc các lớp khác nhau sao cho khoảng cách (margin) giữa siêu phẳng và các điểm dữ liệu gần nhất (gọi là support vectors) là lớn nhất. Khoảng cách này được gọi là margin, và siêu phẳng có margin lớn nhất được gọi là siêu phẳng tối ưu.

- Xử lý dữ liệu không phân tách tuyến tính: Trong trường hợp dữ liệu không thể phân tách tuyến tính, SVC sử dụng kernel functions để ánh xạ dữ liệu vào không gian nhiều chiều hơn, nơi mà dữ liệu có thể phân tách tuyến tính. Các kernel phổ biến bao gồm: linear, polynomial, radial basis function (RBF), và sigmoid.

- Tối ưu hóa: Bài toán tìm siêu phẳng tối ưu được chuyển đổi thành một bài toán tối ưu hóa lồi (convex optimization), thường được giải quyết bằng các phương pháp như Sequential Minimal Optimization (SMO).

- Dự đoán: Sau khi huấn luyện, mô hình SVC có thể dự đoán lớp của một điểm dữ liệu mới bằng cách xác định nó nằm ở phía nào của siêu phẳng.

#### 2.2.4 Logistic Regression

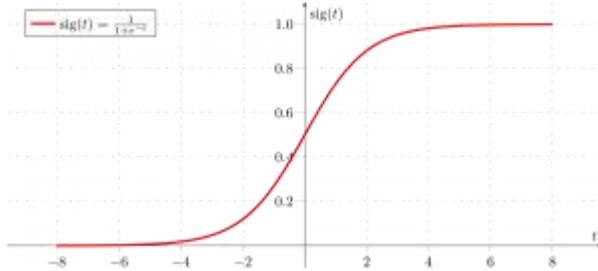
Logistic Regression (Hồi quy Logistic) là một thuật toán Machine Learning dùng để giải quyết bài toán phân loại (Classification). Mặc dù có từ "Regression" (hồi quy), nhưng nó không được dùng để dự đoán giá trị liên tục như Hồi quy tuyến tính (Linear Regression) mà thay vào đó, Logistic Regression dự đoán xác suất của một sự kiện thuộc về một trong hai nhóm (phân loại nhị phân) hoặc nhiều nhóm (phân loại đa lớp).<sup>19</sup>

Logistic Regression thường được sử dụng để phân loại dữ liệu vào hai nhóm. Ví dụ, dự đoán liệu một email có phải là spam hay không, hoặc liệu một bệnh nhân có mắc một bệnh cụ thể hay không.

Logistic Regression sử dụng hàm logistic (còn gọi là hàm sigmoid) để chuyển đổi đầu ra của một phương trình tuyến tính thành một giá trị xác suất nằm trong khoảng từ 0 đến 1.

---

<sup>19</sup> Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). Springer.



Hình 2.3: Hàm Sigmoid<sup>20</sup>

Hồi quy Logistic hoạt động dựa trên hàm Sigmoid, được biểu diễn như sau:

$$S(z) = \frac{1}{(1 + e^{-z})}^{21}$$

Hàm Sigmoid nhận đầu vào là một giá trị  $z$  bất kỳ, và trả về đầu ra là một giá trị xác suất nằm trong khoảng  $[0,1]$ . Khi áp dụng vào mô hình Hồi quy Logistic với đầu vào là ma trận dữ liệu  $X$  và trọng số  $w$ , ta có  $z = Xw$ .

Việc huấn luyện của mô hình là tìm ra bộ trọng số  $w$  sao cho đầu ra dự đoán của hàm Sigmoid gần với kết quả thực tế nhất. Để làm được điều này, ta sử dụng hàm mất mát (Loss Function) để đánh giá hiệu năng của mô hình. Mô hình càng tốt khi hàm mất mát càng nhỏ.

Hàm mất mát (Loss Function) là một hàm số được sử dụng để đo lường mức độ lỗi mà mô hình của chúng ta tạo ra khi dự đoán các kết quả từ dữ liệu đầu vào. Trong bài toán Hồi quy Logistic, chúng ta sử dụng hàm mất mát Cross-Entropy (còn gọi là Log Loss) để đánh giá hiệu năng của mô hình.

Hàm mất mát Cross-Entropy được định nghĩa như sau:

$$L(w) = -\frac{1}{n} \sum_{i=1}^n [y_i \log p_i + (1 - y_i) \log(1 - p_i)]^{22}$$

Trong đó:

$n$ : số lượng mẫu dữ liệu trong tập huấn luyện

$y_i$ : giá trị thực tế của đầu ra thứ  $i$

$p_i$ : xác suất dự đoán thuộc lớp 1 của mô hình cho đầu vào thứ  $i$ .

Hàm Cross-Entropy đo lường khoảng cách giữa hai phân phối xác suất  $y_i$  và  $p_i$ . Khi mô hình dự đoán chính xác, tức là nếu  $y_i = 1$  thì  $p_i$  càng gần 1, và nếu  $y_i = 0$  thì  $p_i$  càng gần 0, sau đó hàm mất mát sẽ tiến gần về 0.

<sup>20</sup> Wikipedia. (n.d.). *Sigmoid function* [Image]. Truy cập từ <https://de.m.wikipedia.org/wiki/Datei:Sigmoid-function-2.svg>

<sup>21</sup> Berkson, J. (1944). Application of the Logistic Function to Bio-Assay. *Journal of the American Statistical Association*.

<sup>22</sup> Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Trong quá trình huấn luyện, chúng ta tìm cách cập nhật bộ trọng số  $w$  sao cho giá trị hàm mất mát Cross-Entropy đạt giá trị nhỏ nhất, dẫn đến một mô hình dự đoán tốt nhất.

Để tìm giá trị tối ưu cho bộ trọng số  $w$ , chúng ta có thể sử dụng kỹ thuật Gradient Descent. Tại mỗi bước lặp, chúng ta cập nhật  $w$  theo phương trình ứng với đạo hàm của hàm mất mát  $L(w)$  theo  $w$ .<sup>23</sup>

### 2.3 Mô hình học sâu LSTM

Long Short-Term Memory Networks thường được gọi là mạng "LSTM", là một loại cụ thể của mạng nơ-ron hồi quy (RNN). Nó được thiết kế để giải quyết vấn đề mất mát gradient trong quá trình học các chuỗi dữ liệu dài hạn<sup>24</sup>.

Ban đầu, ý tưởng chính của mạng RNN là sử dụng chuỗi thông tin. Trong các mạng nơ ron truyền thống thì hầu hết các thông tin độc lập với nhau, nhưng điểm nổi bật của RNN chính là ý tưởng kết nối các thông tin phía trước, vòng lặp bên trong cho phép thông tin có thể lưu lại được từ đó dự đoán cho hiện tại. Chẳng hạn, ta có câu: “các đám mây trên bầu trời” thì ta chỉ cần đọc tới “các đám mây trên bầu” là đủ biết được chữ tiếp theo là “trời” rồi. Trong tình huống này, RNN hoàn toàn có thể dự đoán được từ kế tiếp. Tuy nhiên, RNN thường gặp phải hai vấn đề chính khi xử lý dữ liệu chuỗi dài:

- Gradient biến mất (vanishing gradient): Khi chuỗi dữ liệu đầu vào kéo dài, gradient của các bước thời gian trước có xu hướng nhỏ dần và gần bằng 0, khiến mô hình khó học được mối quan hệ xa trong chuỗi.

- Gradient bùng nổ (exploding gradient): Trong một số trường hợp, gradient có thể tăng đột biến, dẫn đến việc làm mất ổn định mô hình.<sup>25</sup>

Do đó, ta có mạng LSTM khắc phục vấn đề phụ thuộc xa, dựa vào đặc điểm nổi trội là việc nhớ thông tin trong suốt thời gian dài.

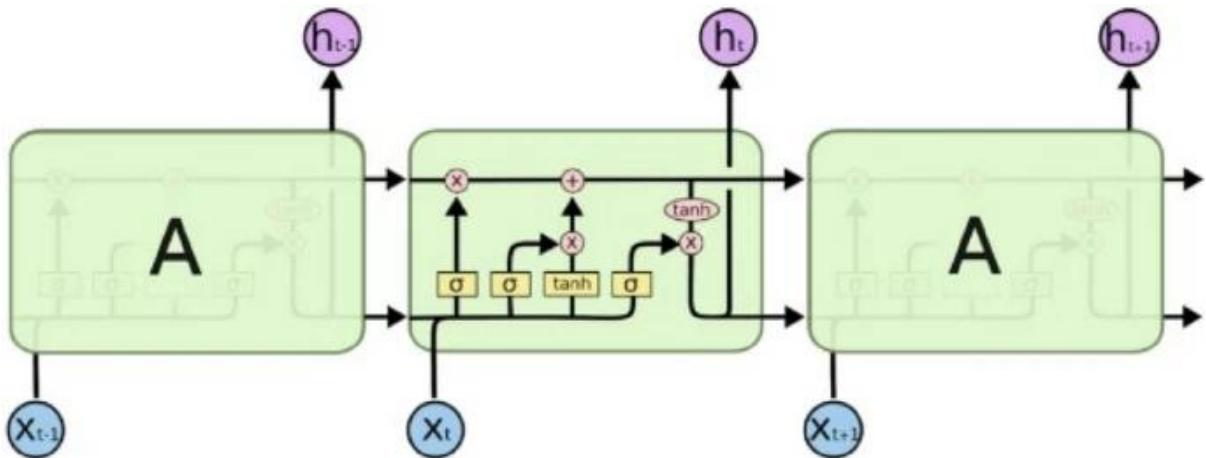
Được biết, LSTM có cấu trúc dạng chuỗi, nhưng các mô-đun có cấu trúc khác so với RNN. Thay vì có lớp nơ ron đơn, LSTM có 4 lớp, chúng tương tác với nhau.

Hình dưới đây mô tả cấu trúc mạng LSTM.

<sup>23</sup> Boyd, S., & Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

<sup>24</sup> Hochreiter, S., & Schmidhuber, J. (1997). *Long Short-Term Memory*. *Neural Computation*, 9(8), 1735–1780.

<sup>25</sup> Hochreiter, S., & Schmidhuber, J. (1997). *Long Short-Term Memory*. *Neural Computation*.



Hình 2.4: Mô tả cấu trúc mạng LSTM và cấu trúc của một mô-đun của nó<sup>26</sup>

Ý tưởng cốt lõi của LSTM là trạng thái của các tế bào được mô tả bằng đường thẳng nằm ngang ở trên cùng. Trạng thái tế bào giống như một băng chuyên. Nó chạy thẳng xuyên qua toàn bộ chuỗi, chỉ với một số tương tác tuyến tính nhỏ, nó rất dễ dàng để thông tin trôi theo dòng không thay đổi.

Một đơn vị LSTM cơ bản bao gồm một ô bộ nhớ, một cổng đầu vào (Input Gate), một cổng đầu ra (Output Gate) và một cổng quên (Forget Gate). Ban đầu, cổng quên không phải là một phần của mạng LSTM mà được đề xuất bởi Gers và cộng sự<sup>27</sup> để cho phép mạng đặt lại trạng thái của nó. Ô bộ nhớ có nhiệm vụ ghi nhớ các giá trị trong các khoảng thời gian tùy ý, trong khi ba cổng còn lại điều chỉnh luồng thông tin liên quan đến ô bộ nhớ. Trong phần còn lại của mục này, thuật ngữ LSTM sẽ chỉ phiên bản cơ bản, vì đây là kiến trúc LSTM phổ biến nhất<sup>28</sup>. Tuy nhiên, điều này không có nghĩa là nó luôn là lựa chọn tối ưu trong mọi tình huống.

Tiếp theo, nhằm hiểu rõ cách thức hoạt động của một mô hình LSTM, quá trình này được mô tả cụ thể bên dưới được đề xuất bởi.

Bước đầu tiên trong LSTM là quyết định thông tin nào sẽ bị loại bỏ khỏi trạng thái tế bào, được thực hiện bởi một lớp *sigmoid* được gọi là “lớp cổng quên”. Đầu vào của nó là  $h_{t-1}$  và  $x_t$ , và cho ra một giá trị thuộc đoạn  $[0, 1]$  cho mỗi trạng thái. Nếu

<sup>26</sup> S. Hochreiter, and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.

<sup>27</sup> Gers, F., Schmidhuber, J., Cummins, F.: Learning to forget: Continual pre-diction with lstm. Neural computation 12, 2451-71 (2000).

<sup>28</sup> He, X., Shi, B., Bai, X., Xia, G.S., Zhang, Z., Dong, W.: Image caption generation with part of speech guidance. Pattern Recognition Letters 119, 229 – 237 (2019). Deep Learning for Pattern Recognition

giá trị là 1 thì các thông tin được giữ lại hoàn toàn, nếu giá trị là 0 có nghĩa là các thông tin bị loại bỏ hoàn toàn.

$$f_t = (\dots [h_{t-1}, x_t] + b_f) \quad (1)^{29}$$

Bước tiếp theo là quyết định thông tin mới sẽ được lưu trữ trong trạng thái tế bào. Để thực hiện việc này, chúng phải thực hiện bằng 2 lớp, một lớp sigmoid được gọi là "lớp cổng đầu vào" quyết định những giá trị nào sẽ được cập nhật. Sau đó, một lớp *tanh* tạo ra một vectơ mới,  $C_t$ , có thể được thêm vào trạng thái.

Tiếp theo, chúng ta kết hợp hai thành phần này để tạo bản cập nhật cho trạng thái.

$$\begin{aligned} &= (\dots [h_{t-1}, x_t] + b_t) \quad (2) \\ C'_t &= \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (3)^{30} \end{aligned}$$

Bây giờ, chúng ta cập nhật trạng thái tế bào cũ, vào trạng thái tế bào mới  $C_t$  như sau:

$$C_t = f_t * -1 + * C'_t \quad (4)^{31}$$

Cuối cùng, tính giá trị đầu ra dựa trên trạng thái tế bào nhưng nó là một phiên bản đã được lọc. Trước tiên, chúng ta thực hiện lớp sigmoid để quyết định phần trạng thái tế bào sẽ xuất ra, sau đó đẩy trạng thái tế bào qua *tanh* và nhân nó với đầu ra của cổng sigmoid.

$$\begin{aligned} o_t &= (\dots [h_{t-1}, x_t] + b_o) \quad (6) \\ h_t &= o_t * \tanh(C_t) \quad (7)^{32} \end{aligned}$$

Hiện tại LSTM đã được chứng minh hoạt động hiệu quả trong việc xử lý dữ liệu qua nhiều nghiên cứu đã được chứng minh trước đó. Có thể nói đến Graves và cộng sự<sup>33</sup> đã áp dụng LSTM vào nhận dạng chữ viết và tổng hợp văn bản, cho thấy mô hình này có khả năng nắm bắt các mối quan hệ lâu dài trong dữ liệu tuần tự. Ngoài ra, trong bài toán dự báo chuỗi thời gian, Qin và cộng sự<sup>34</sup> đã đề xuất mô hình LSTM kết hợp với cơ chế attention, giúp cải thiện độ chính xác đáng kể so với các phương pháp

<sup>29</sup> Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735-1780.

<sup>30</sup> Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735-1780.

<sup>31</sup> Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735-1780.

<sup>32</sup> Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735-1780.

<sup>33</sup> Graves, A., Mohamed, A. R., & Hinton, G. (2013). *Speech recognition with deep recurrent neural networks*. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing.

<sup>34</sup> Qin, Y., Song, D., Chen, H., Cheng, W., Jiang, G., & Cottrell, G. (2017). A dual-stage attention-based recurrent neural network for time series prediction. *Advances in Neural Information Processing Systems*.

truyền thống. Từ những nghiên cứu này khẳng định rằng LSTM là một trong những mô hình hiệu quả nhất để xử lý dữ liệu chuỗi, mở ra nhiều hướng ứng dụng trong khoa học và công nghệ.

## CHƯƠNG 3: PHƯƠNG PHÁP ĐỀ XUẤT

### 3.1 Thu thập dữ liệu

Để tập trung vào việc phân tích nội dung văn bản thông qua thuật toán Sentiment Analysis, nghiên cứu này sử dụng dữ liệu từ Agoda làm nguồn chính. Agoda là một nền tảng đặt phòng trực tuyến phổ biến, cung cấp hàng triệu đánh giá từ người dùng về khách sạn, khu nghỉ dưỡng và các dịch vụ lưu trú khác. Nhờ đó, dữ liệu từ Agoda có thể phản ánh chân thực trải nghiệm của khách du lịch và hỗ trợ hiệu quả cho việc đánh giá cảm xúc trong nhận xét của họ.

Tuy nhiên, để huấn luyện mô hình dự đoán cảm xúc, nghiên cứu sử dụng tập dữ liệu từ TripAdvisor. TripAdvisor, được thành lập vào năm 2000 bởi Stephen Kaufer và các cộng sự, là một trong những trang web đánh giá du lịch lớn nhất thế giới. Với lượng dữ liệu phong phú và hệ thống đánh giá chi tiết, tập dữ liệu từ TripAdvisor giúp xây dựng một mô hình phân tích cảm xúc chính xác hơn trước khi áp dụng vào dữ liệu thực tế từ Agoda.

Nhằm thu thập dữ liệu huấn luyện, một trình thu thập dữ liệu web đã được phát triển để lấy các bài đánh giá về khách sạn trên TripAdvisor. Sau khi mô hình được huấn luyện và tối ưu hóa trên tập này, nó sẽ được triển khai để phân tích các đánh giá thu thập từ Agoda, từ đó cung cấp những thông tin giá trị về xu hướng cảm xúc của khách hàng đối với các khách sạn trên nền tảng này.

#### 3.1.1 Tập dữ liệu

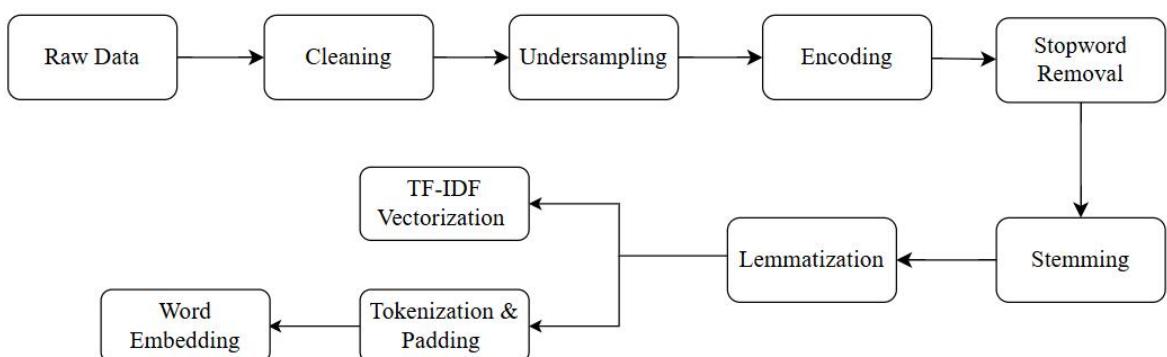
Để đảm bảo có một tập dữ liệu huấn luyện mô hình toàn diện về các đánh giá khách sạn ở Marrakech, nghiên cứu này sử dụng tập dữ liệu từ TripAdvisor được thu thập bởi một nguồn khác, thay vì tự động thu thập qua web scraping. TripAdvisor là một nền tảng du lịch trực tuyến phổ biến, nơi người dùng có thể chia sẻ đánh giá về khách sạn, nhà hàng và các dịch vụ lưu trú khác.

Tập dữ liệu này đã được thu thập và tổng hợp trước đó, giúp tiết kiệm thời gian và đảm bảo tính nhất quán trong dữ liệu đầu vào. Các đánh giá trong tập dữ liệu bao gồm thông tin về khách sạn, nội dung nhận xét của khách hàng và xếp hạng cảm xúc. Nguồn dữ liệu cụ thể được lấy từ: **[marrakech\_hotels\_reviews.csv]**, cung cấp các đánh giá khách sạn tại Marrakech và phục vụ quá trình huấn luyện mô hình Sentiment Analysis.

Việc sử dụng dữ liệu từ TripAdvisor chỉ nhằm mục đích huấn luyện mô hình, trong khi tập dữ liệu thực tế dùng để đánh giá và ứng dụng mô hình sẽ được thu thập từ nền tảng Agoda.

### 3.1.2 Tiền xử lý dữ liệu

Đầu tiên, chuẩn bị dữ liệu là bước đầu tiên trong quá trình phân tích Sentiment Analysis nhằm đảm bảo chất lượng, độ tin cậy và tính phù hợp cho dữ liệu. Quá trình này giúp khám phá những insight có ý nghĩa, cải thiện hiệu suất mô hình và hỗ trợ đưa ra các quyết định sáng suốt dựa trên kết quả Phân tích cảm xúc từ các bài đánh giá khách sạn.



Hình 3.1: Các bước tiền xử lý dữ liệu

#### 3.1.2.1 NLTK and SpaCy

NLTK và SpaCy là hai thư viện Python phổ biến dành cho các tác vụ xử lý ngôn ngữ tự nhiên (NLP). NLTK cung cấp một bộ công cụ toàn diện, cho phép kiểm soát chi tiết quy trình xử lý NLP. Thư viện này bao gồm nhiều chức năng như tách từ (tokenization), rút gọn từ (stemming), gán nhãn từ loại (POS tagging) và phân tích cú pháp (syntactic parsing). Điểm mạnh của NLTK nằm ở tính linh hoạt và nguồn tài nguyên phong phú.

Mặt khác, SpaCy tập trung vào hiệu suất và sự dễ sử dụng. Nó đặc biệt hiệu quả trong việc xử lý dữ liệu quy mô lớn và các ứng dụng thời gian thực. SpaCy cung cấp các chức năng như tách từ, gán nhãn từ loại, phân tích phụ thuộc cú pháp (dependency parsing), nhận diện thực thể có tên (named entity recognition - NER) và liên kết thực thể (entity linking). Thư viện này ưu tiên tốc độ và hiệu suất với một quy trình xử lý tối ưu.

Cả hai thư viện đều có các mô hình được huấn luyện sẵn cho nhiều tác vụ khác nhau. Việc lựa chọn giữa NLTK và SpaCy phụ thuộc vào yêu cầu cụ thể của dự án, trong đó NLTK phù hợp cho nghiên cứu và thử nghiệm, còn SpaCy thích hợp cho các ứng dụng sản xuất quy mô lớn.

### 3.1.2.2 Undersampling

Undersampling là một nhóm các kỹ thuật được thiết kế để cân bằng phân phối lớp trong tập dữ liệu phân loại có sự chênh lệch giữa các lớp. Trong một tập dữ liệu mất cân bằng, sẽ có một hoặc nhiều lớp có ít mẫu (lớp thiểu số) và một hoặc nhiều lớp có nhiều mẫu (lớp đa số). Và Undersampling bao gồm việc giảm dữ liệu bằng cách loại bỏ các mẫu thuộc lớp đa số nhằm mục đích cân bằng số lượng mẫu giữa các lớp (Trang 82, Learning from Imbalanced Data Sets, 2018). Trong mô hình này, nhóm chúng tôi đã sử dụng random undersampling, tức là chọn ngẫu nhiên một số lượng mẫu nhất định từ các nhóm có Review\_Rating là 5,4,3 và loại bỏ một số lượng mẫu nhất định từ các nhóm này nhằm cân bằng lại tập dữ liệu.

```
remove_indices_5 = np.random.choice(df[mask_5].index, size=32000, replace=False)  
remove_indices_4 = np.random.choice(df[mask_4].index, size=9156, replace=False)  
remove_indices_3 = np.random.choice(df[mask_3].index, size=3000, replace=False)
```

### 3.1.2.3 Encoding

Encoding là quá trình chuyển dữ liệu từ một định dạng có sẵn, thường là từ dữ liệu phi số (như văn bản, hình ảnh, danh mục) sang dạng số để có thể sử dụng trong các mô hình máy học. Có nhiều phương pháp encoding phổ biến, nhưng trong nghiên cứu này tập trung vào Label encoding, trong đó dữ liệu dạng danh mục được chuyển thành số nguyên.

"Good" → 1 "Bad" → 0

### 3.1.2.4 Stopword Removal

Stopword Removal là một kỹ thuật tiền xử lý phổ biến trong Xử lý Ngôn ngữ Tự nhiên (NLP), trong đó các từ xuất hiện thường xuyên nhưng không mang ý nghĩa quan trọng cho nhiệm vụ cần thực hiện sẽ bị loại bỏ. Stopwords là những từ như "the", "is",

"and" và "in", xuất hiện thường xuyên trong một ngôn ngữ nhưng ít đóng góp vào việc hiểu nội dung tổng thể của văn bản.<sup>35</sup>

```
clean_text = [word for word in clean_text.split() if word not in topwords.words('english')]
```

Với cú pháp trên, mô hình giúp giảm kích thước dữ liệu văn bản, điều này đặc biệt hữu ích khi làm việc với các tập dữ liệu lớn. Bằng cách loại bỏ stopwords, quá trình phân tích có thể tập trung vào những từ mang ý nghĩa quan trọng hơn, giúp nâng cao hiệu suất và hiệu quả xử lý. Thứ hai, việc loại bỏ stopwords có thể cải thiện độ chính xác và hiệu suất của nhiều tác vụ NLP, chẳng hạn như phân loại văn bản, phân tích cảm xúc và truy xuất thông tin.

This hotel is my absolute favorite.

↓  
Stopword removal

This hotel is my absolute favorite.

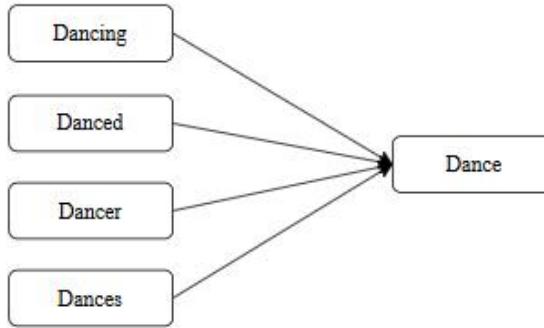
Hình 3.2: Minh họa việc loại bỏ stopwords

### 3.1.2.5 Stemming

Stemming là một kỹ thuật tiền xử lý văn bản trong xử lý ngôn ngữ tự nhiên (NLP). Cụ thể, đây là quá trình đưa các dạng biến tố của một từ về một dạng gốc, được gọi là "stem".<sup>36</sup> Mục tiêu của Stemming là chuẩn hóa từ bằng cách loại bỏ các tiền tố hoặc hậu tố. Quá trình này giúp hợp nhất các biến thể khác nhau cùng một từ, tạo điều kiện cho việc phân tích và so sánh dữ liệu văn bản hiệu quả hơn.

<sup>35</sup> Removing stop words with NLTK in Python (January 03, 2024), <https://www.geeksforgeeks.org/removing-stop-words-nltk-python/>

<sup>36</sup> Ruslan Mitkov, Oxford Handbook of Computational Linguistics, 2nd edition, Oxford University Press, 2014.



Hình 3.3: Mô tả Stemming

### 3.1.2.6 Lemmatization

Lemmatization là một quá trình ngôn ngữ học thường được sử dụng trong xử lý ngôn ngữ tự nhiên (NLP) để đưa các từ về dạng cơ bản hoặc dạng từ điển của chúng, được gọi là lemma. Khác với stemming, vốn chỉ tập trung vào việc loại bỏ tiền tố hoặc hậu tố, lemmatization thực hiện phân tích hình thái của từ và xem xét ngữ cảnh cũng như loại từ (POS) của từng từ. Bằng cách áp dụng lemmatization, các từ được chuyển đổi về dạng chuẩn của chúng, giúp giảm bớt các từ bị biến đổi hoặc phát sinh về cùng một gốc chung.

```

lemmatizer = WordNetLemmatizer()
sentence.append(lemmatizer.lemmatize(word, 'v'))

```

Ở đây, với cú pháp WordNetLemmatizer() giúp chuyển đổi các từ về dạng cơ bản nhất, chẳng hạn:

running → run; studies → study; better → good

### 3.1.2.7 TF-IDF Vectorization

TF-IDF Vectorization là một kỹ thuật trong Xử lý Ngôn ngữ Tự nhiên (NLP) được sử dụng để chuyển đổi văn bản thành dạng số, giúp máy học hiểu và xử lý dữ liệu văn bản. Nó đánh giá mức độ quan trọng của một từ trong một tài liệu dựa trên tần suất xuất hiện của từ đó trong tài liệu (TF) và mức độ phổ biến của từ trong toàn bộ tập dữ liệu (IDF). Các từ xuất hiện nhiều nhưng ít ý nghĩa (như "the", "is") sẽ có trọng số thấp, trong khi các từ quan trọng nhưng ít xuất hiện sẽ có trọng số cao hơn. TF-IDF thường được sử dụng trong phân loại văn bản, tìm kiếm thông tin và phân tích cảm xúc (Rajaraman & Ullman, 2011). Trong mô hình này, TF-IDF được dùng để

biến đổi dữ liệu thành dạng số trước khi đưa vào làm việc với các mô hình học máy truyền thống như BernoulliNB, Logistic Regression, Random Forest và SVC.

### 3.1.2.8 Tokenization & Padding

Tokenization & Padding là hai bước quan trọng trong xử lý ngôn ngữ tự nhiên (NLP), đặc biệt khi làm việc với mô hình học sâu như LSTM. Với Tokenization, là quá trình chuyển đổi văn bản thô thành một danh sách các token (từ hoặc ký tự). Chẳng hạn:

*Văn bản:* "Hôm nay trời đẹp quá!"

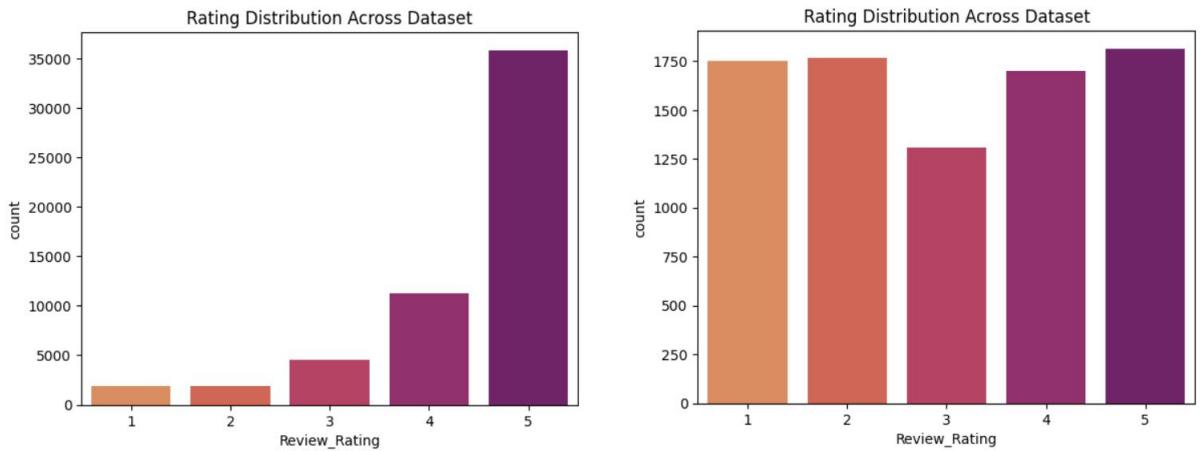
*Sau khi tokenization (word-level):* ["Hôm", "nay", "trời", "đẹp", "quá", "!"].

Trong mô hình học sâu, mỗi token sẽ được ánh xạ thành một chỉ số số nguyên (index) dựa trên từ điển được xây dựng từ tập dữ liệu. Còn Padding giúp các mô hình như LSTM xử lý dữ liệu dễ dàng hơn, tránh lỗi khi đầu vào có độ dài không đồng nhất do các mô hình như LSTM yêu cầu đầu vào có độ dài cố định.

### 3.1.2.9 Word Embedding

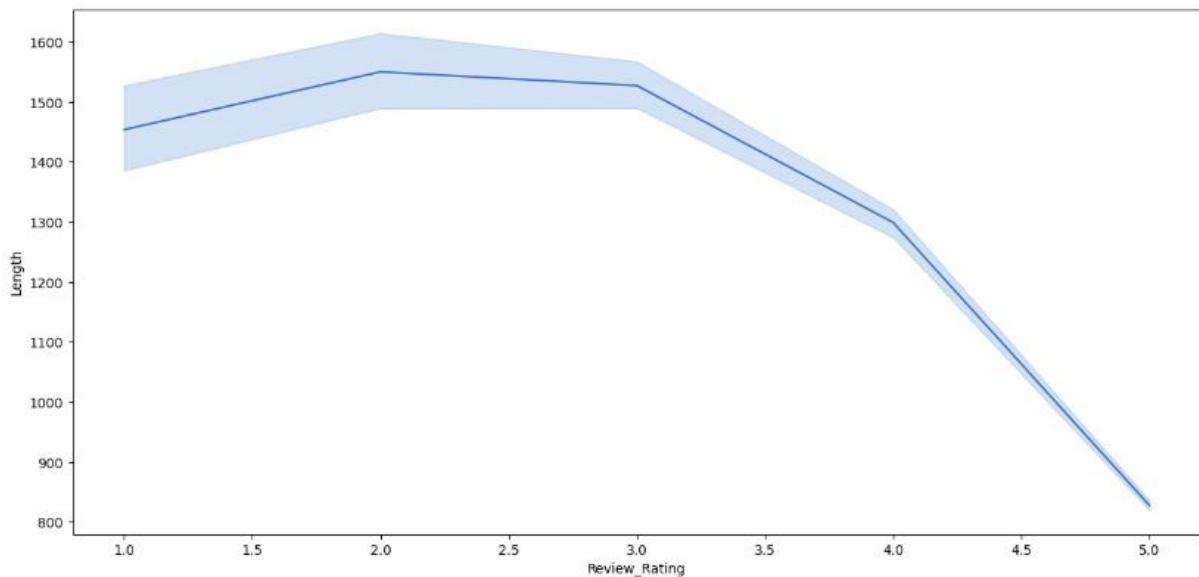
Word Embedding là một không gian vector dùng để biểu diễn dữ liệu có khả năng miêu tả được mối liên hệ, sự tương đồng về mặt ngữ nghĩa, văn cảnh(context) của dữ liệu. Thay vì biểu diễn từ bằng các chỉ số đơn giản (one-hot encoding), Word Embedding giúp mã hóa mối quan hệ giữa các từ bằng cách đặt chúng vào không gian nhiều chiều, nơi các từ có ý nghĩa tương tự nằm gần nhau hơn. Ví dụ như ta có hai câu : "Hôm nay ăn táo " và "Hôm nay ăn xoài ". Khi ta thực hiện Word Embedding, "táo" và "xoài" sẽ có vị trí gần nhau trong không gian chúng ta biểu diễn do chúng có vị trí giống nhau trong một câu. Với mô hình nhóm xây dựng, Word Embedding được sử dụng sau bước Tokenization & Padding để chuyển các token thành vector số trước khi đưa vào mô hình LSTM. Nhờ vậy, mô hình có thể học được ý nghĩa ngữ nghĩa và mối quan hệ giữa các từ, giúp tăng độ chính xác.

### 3.1.3 Trực quan hóa dữ liệu



Hình 3.4: Tổng quan phân bố lượt đánh giá

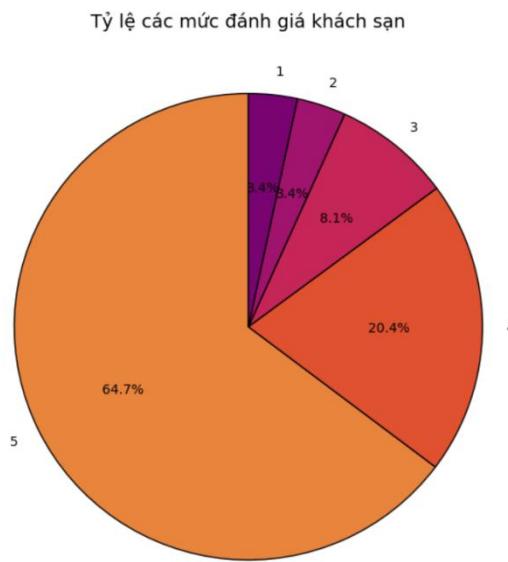
Với biểu đồ thể hiện tổng quan phân bố lượt đánh giá trước khi áp dụng kỹ thuật UnderSampling, nhận thấy rằng có sự chênh lệch lớn về số lượng đánh giá giữa các mức rating. Đánh giá 5 sao chiếm phần lớn, trong khi các mức thấp hơn (1, 2, 3 sao) có số lượng rất ít. Điều này phản ánh sự khác biệt đáng kể trong trải nghiệm của khách hàng, có thể dẫn đến sai lệch trong phân tích dữ liệu. Sau khi áp dụng kỹ thuật sampling để cân bằng dữ liệu, số lượng đánh giá ở các mức rating trở nên đồng đều hơn. Việc này giúp giảm thiểu sự mất cân bằng trong dữ liệu, từ đó cải thiện độ chính xác của các mô hình học máy và giảm bias trong quá trình phân tích.



Hình 3.5: Tổng quan độ dài theo mức độ đánh giá

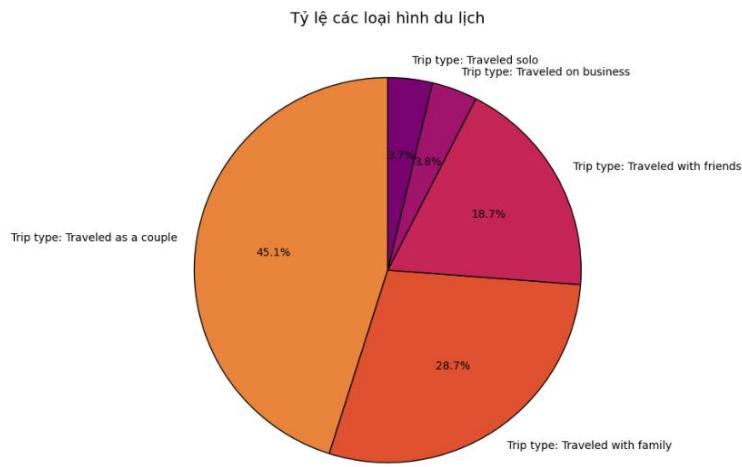
Dễ dàng nhận thấy rằng các đánh giá có mức rating thấp (1,2 và 3 sao) thường có độ dài trung bình cao hơn so với các mức rating cao hơn. Khi mức đánh giá tăng lên,

đặc biệt từ 3 sao trở đi, độ dài trung bình của đánh giá có xu hướng giảm dần, trong đó các đánh giá 5 sao thường ngắn nhất. Phần dải màu xanh thể hiện mức độ biến động của dữ liệu, cho thấy dù có sự dao động nhưng vẫn tuân theo xu hướng chung. Điều này phản ánh một thực tế phổ biến: người dùng có trải nghiệm tiêu cực thường để lại nhận xét chi tiết hơn, trong khi những khách hàng hài lòng thường chỉ đưa ra phản hồi ngắn gọn.



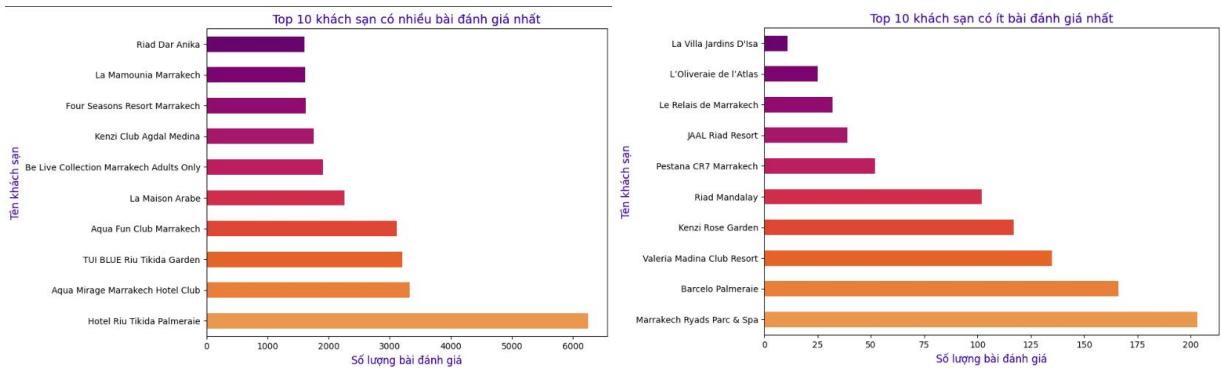
Hình 3.6: Tỷ lệ các mức đánh giá khách sạn

Phân bố đánh giá trong tập dữ liệu đánh giá khách sạn tiết lộ rằng phần lớn khách hàng đã thể hiện cảm xúc rất tích cực, thể hiện qua số lượng lớn các đánh giá ở mức cao nhất trên thang điểm, cụ thể là 4 và 5. Điều này cho thấy một tỷ lệ đáng kể khách hàng đã có trải nghiệm xuất sắc và hài lòng cao với các khách sạn. Tuy nhiên, cũng cần lưu ý sự xuất hiện của các đánh giá thấp hơn, đặc biệt là 1 và 2, cho thấy những trường hợp có cảm xúc tiêu cực và sự không hài lòng. Mặc dù các đánh giá thấp này chiếm tỷ lệ nhỏ trong tổng thể, nhưng chúng vẫn phản ánh những khía cạnh cần cải thiện để giải quyết mối quan tâm của khách hàng và nâng cao trải nghiệm của họ. Sự phân bố này thể hiện phạm vi cảm xúc mà khách hàng bày tỏ trong các bài đánh giá khách sạn, nhấn mạnh tầm quan trọng của việc hiểu cả những điểm tích cực lẫn những lĩnh vực cần cải thiện để có cái nhìn toàn diện về mức độ hài lòng của khách hàng.



*Hình 3.7: Tỷ lệ các loại hình du lịch*

Phân bố loại hình du lịch trong tập dữ liệu cung cấp những thông tin quan trọng về nhân khẩu học và sở thích của người đánh giá. Trong số các loại hình chuyến đi, danh mục phổ biến nhất được quan sát là loại hình du lịch "cặp đôi", cho thấy một số lượng đáng kể du khách đã đến khách sạn để tận hưởng những kỳ nghỉ lãng mạn. Sau "cặp đôi", loại hình chuyến đi phổ biến tiếp theo là "gia đình", cho thấy một phần đáng kể những người đánh giá là các gia đình tìm kiếm các lựa chọn lưu trú phù hợp với nhu cầu của họ. Ngoài ra, danh mục "bạn bè" cũng cho thấy một số lượng lớn người đánh giá đã đi du lịch cùng bạn bè, có thể nhằm mục đích trải nghiệm chung và tham gia các hoạt động nhóm. Đáng chú ý, các loại hình chuyến đi "công tác" và "du lịch một mình" có số lượng thấp hơn đáng kể so với các danh mục khác, cho thấy tập dữ liệu chủ yếu bao gồm các đánh giá liên quan đến du lịch nghỉ dưỡng. Việc hiểu được sự phân bố của các loại hình chuyến đi rất quan trọng đối với các khách sạn trong việc điều chỉnh dịch vụ và tiện nghi nhằm đáp ứng nhu cầu và sở thích đa dạng của đối tượng khách hàng mục tiêu.

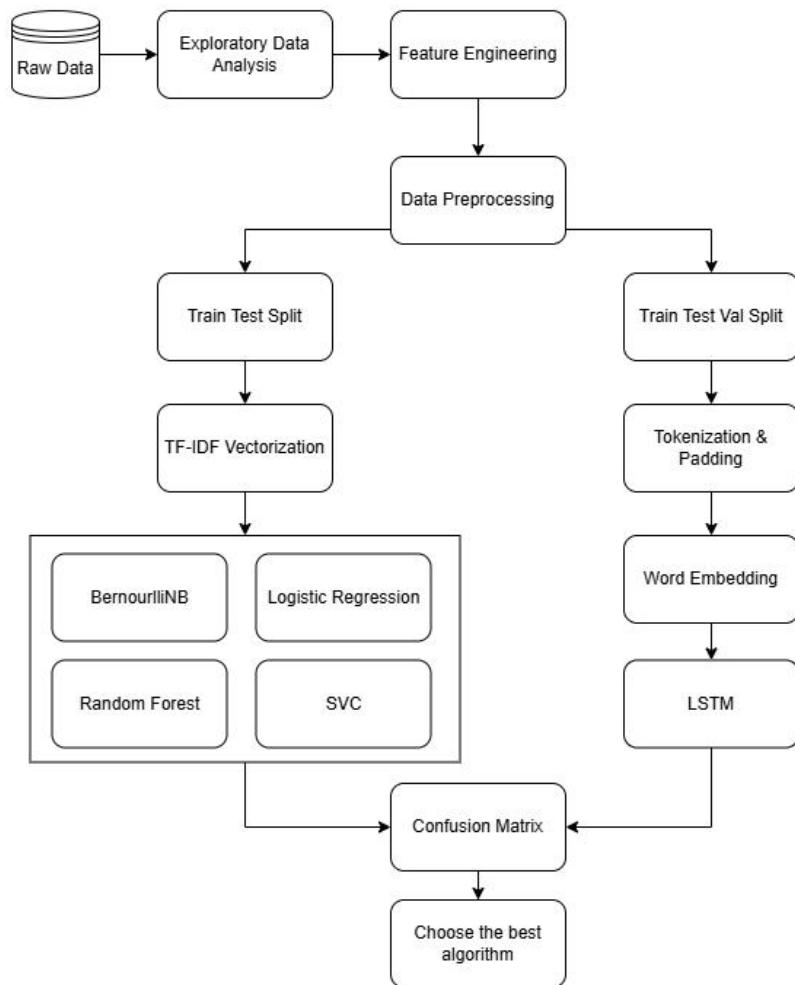


Hình 3.8: Xếp hạng số lượng bài đánh giá của các khách sạn

Biểu đồ bên trái thể hiện danh sách 10 khách sạn có nhiều bài đánh giá nhất, trong đó trực hoành biểu diễn số lượng bài đánh giá, còn trực tung hiển thị tên khách sạn. Các thanh màu từ tím đến cam thể hiện mức độ đánh giá, với khách sạn có nhiều bài đánh giá nhất có thanh dài nhất. Ngược lại, biểu đồ bên phải hiển thị 10 khách sạn có ít bài đánh giá nhất, sử dụng cùng cách trình bày nhưng với số lượng bài đánh giá thấp hơn. Cả hai biểu đồ giúp người xem so sánh mức độ quan tâm của khách hàng đối với các khách sạn dựa trên số lượng bài đánh giá được ghi nhận.

### 3.2 Quy trình xây dựng mô hình NLP

Pipeline của mô hình NLP bao gồm nhiều bước xử lý từ dữ liệu thô đến đánh giá mô hình và chọn ra thuật toán phù hợp nhất. Mục tiêu chính là chọn ra thuật toán tối ưu nhất để có thể phân tích cảm xúc của bình luận là tích cực hay tiêu cực, từ đó đưa ra các quyết định dựa trên dữ liệu một cách chính xác và hiệu quả.



Hình 3.9: Quy trình xây dựng mô hình NLP

Quá trình bắt đầu với việc thu thập dữ liệu thô. Dữ liệu này thường bao gồm văn bản và nhãn cảm xúc (tích cực hoặc tiêu cực). Sau khi thu thập, dữ liệu được đưa vào giai đoạn phân tích dữ liệu khám phá (Exploratory Data Analysis - EDA) để hiểu rõ về đặc điểm của dữ liệu, bao gồm việc kiểm tra tỷ lệ bình luận tích cực và tiêu cực, tìm kiếm giá trị bị thiếu, kiểm tra sự xuất hiện của các từ phổ biến và phát hiện nhiều trong dữ liệu. Tiếp theo, quá trình trích xuất đặc trưng (Feature Engineering) được thực hiện. Sau đó, dữ liệu đi qua bước tiền xử lý (Data Preprocessing), nơi các bình luận được làm sạch và chuẩn hóa hoàn toàn. Các bước này bao gồm chuẩn hóa văn bản như chuyển chữ hoa thành chữ thường, loại bỏ dấu câu, loại bỏ stopwords (các từ

không mang nhiều ý nghĩa như "và", "hoặc") và có thể cả stemming hoặc lemmatization để giảm số lượng từ đồng nghĩa. Mục tiêu của bước này là chuyển đổi dữ liệu văn bản thành định dạng phù hợp để đưa vào mô hình huấn luyện.

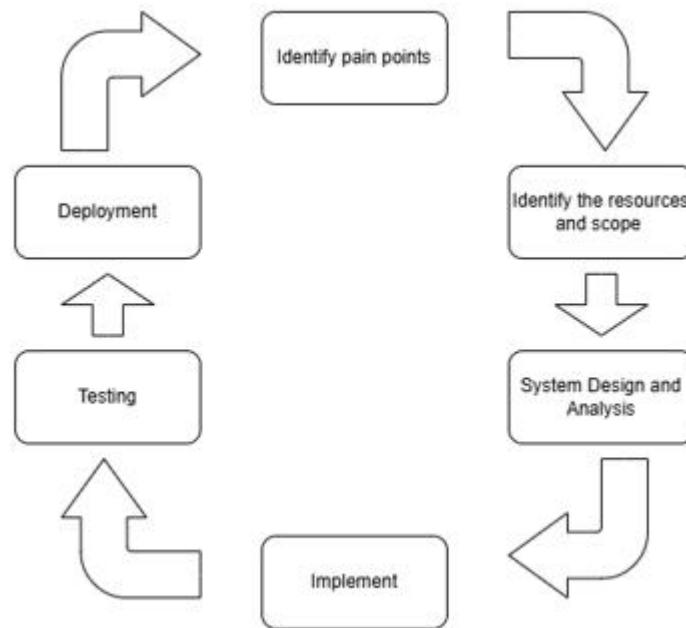
Tại đây, pipeline được chia thành hai hướng tiếp cận:

*Hướng thứ nhất* là sử dụng học máy truyền thống. Dữ liệu sau tiền xử lý sẽ được chia thành tập huấn luyện và tập kiểm tra (Train-Test Split). Tiếp theo, văn bản được chuyển đổi thành dạng số bằng phương pháp TF-IDF (Term Frequency-Inverse Document Frequency) Vectorization, giúp biểu diễn mức độ quan trọng của từng từ trong văn bản. Sau khi có dữ liệu số hóa, các mô hình học máy khác nhau như Bernoulli Naïve Bayes, Random Forest, Logistic Regression, và Support Vector Classifier (SVC) được sử dụng để huấn luyện và dự đoán cảm xúc. Kết quả dự đoán được đánh giá bằng Ma trận nhầm lẫn (Confusion Matrix) để đo lường độ chính xác, độ nhạy và độ đặc hiệu của từng mô hình.

*Hướng thứ hai* là sử dụng học sâu với mô hình LSTM. Sau bước tiền xử lý, dữ liệu được chia thành tập huấn luyện, tập kiểm tra và tập validation (Train-Test-Validation Split). Tiếp theo, dữ liệu văn bản được tokenization để chuyển đổi các từ thành các số nguyên tương ứng và padding để đảm bảo tất cả các câu có cùng độ dài. Sau đó, các token này được đưa qua Word Embedding, một kỹ thuật ánh xạ từ vựng sang không gian vectơ nhiều chiều, giúp mô hình hiểu được ngữ nghĩa của các từ. Cuối cùng, dữ liệu đi vào mô hình LSTM (Long Short-Term Memory), một dạng mạng nơ-ron hồi quy mạnh mẽ trong xử lý ngôn ngữ tự nhiên. Kết quả của mô hình cũng được đánh giá bằng Ma trận nhầm lẫn (Confusion Matrix).

Sau khi đã huấn luyện và đánh giá các mô hình theo cả hai hướng, chúng ta so sánh kết quả từ confusion matrix của từng mô hình. Mô hình nào có hiệu suất tốt nhất dựa trên ma trận nhầm lẫn sẽ được lựa chọn làm thuật toán tối ưu để phân tích cảm xúc bình luận.

### 3.3. Kiến trúc ứng dụng và luồng UI/UX



Hình 3.10: Quy trình phát triển ứng dụng

Quá trình bắt đầu bằng việc xác định các điểm khó khăn (Identify Pain Points). Đây là bước đầu tiên và quan trọng trong việc phát triển một ứng dụng vì nó đặt nền móng cho toàn bộ dự án. Việc xác định các điểm khó khăn giúp đảm bảo rằng ứng dụng sẽ giải quyết đúng nhu cầu của người dùng thay vì chỉ tạo ra một sản phẩm không có tính ứng dụng cao. Ở giai đoạn này, có thể sẽ có các cuộc khảo sát, thu thập phản hồi từ người dùng, hoặc phân tích thị trường để tìm ra những khó khăn mà ứng dụng có thể giải quyết. Mục tiêu là để hiểu rõ những gì cần được cải thiện hoặc giải quyết, từ đó định hướng cho các bước tiếp theo.

Sau khi xác định được các vấn đề cần giải quyết, nhóm sẽ xác định tài nguyên và phạm vi dự án (Identify the resources and scope). Đây là bước đánh giá nguồn lực hiện có, bao gồm nhân lực, công nghệ, thời gian, và ngân sách để hoàn thành dự án. Đồng thời, phạm vi của dự án cũng được xác định để tránh việc mở rộng quá mức hoặc thay đổi liên tục trong quá trình phát triển, điều này có thể gây ra sự chậm trễ và tốn kém không cần thiết. Việc này giúp đảm bảo rằng dự án sẽ được thực hiện một cách khả thi và hiệu quả, đồng thời tránh được những rủi ro không đáng có.

Tiếp theo là giai đoạn thiết kế hệ thống và phân tích (System Design and Analysis). Ở bước này, nhóm sẽ lên kế hoạch chi tiết về kiến trúc của ứng dụng. Điều

này bao gồm việc quyết định các thành phần chính của ứng dụng, cách chúng tương tác với nhau, và lựa chọn công nghệ phù hợp. Đây là bước quan trọng để đảm bảo ứng dụng có một nền tảng vững chắc, đảm bảo rằng hệ thống sẽ hoạt động hiệu quả và đáp ứng được các yêu cầu đã đề ra. Bên cạnh đó, việc phân tích hệ thống giúp phát hiện các rủi ro tiềm ẩn trong quá trình phát triển và triển khai.

Sau khi có thiết kế hệ thống hoàn chỉnh, ứng dụng sẽ được triển khai giai đoạn lập trình (Implement). Đây là lúc bắt đầu viết mã nguồn, tích hợp các thành phần và xây dựng các tính năng của ứng dụng dựa trên thiết kế trước đó. Quá trình này đòi hỏi sự phối hợp chặt chẽ để đảm bảo rằng mọi thứ được thực hiện đúng theo thiết kế ban đầu. Đây cũng là lúc mà các ý tưởng bắt đầu trở thành hiện thực.

Khi ứng dụng đã được lập trình xong, nó sẽ bước vào giai đoạn kiểm thử (Testing). Ở bước này, ứng dụng sẽ được kiểm tra kỹ lưỡng để phát hiện lỗi (bugs), đánh giá hiệu suất, đảm bảo trải nghiệm người dùng tốt và đảm bảo rằng các tính năng hoạt động đúng như mong đợi. Việc kiểm thử có thể bao gồm kiểm thử chức năng (functional testing), kiểm thử hiệu năng (performance testing), kiểm thử bảo mật (security testing), và kiểm thử trải nghiệm người dùng (UX testing). Nếu phát hiện ra lỗi, cần quay lại giai đoạn lập trình để sửa chữa và cải thiện ứng dụng.

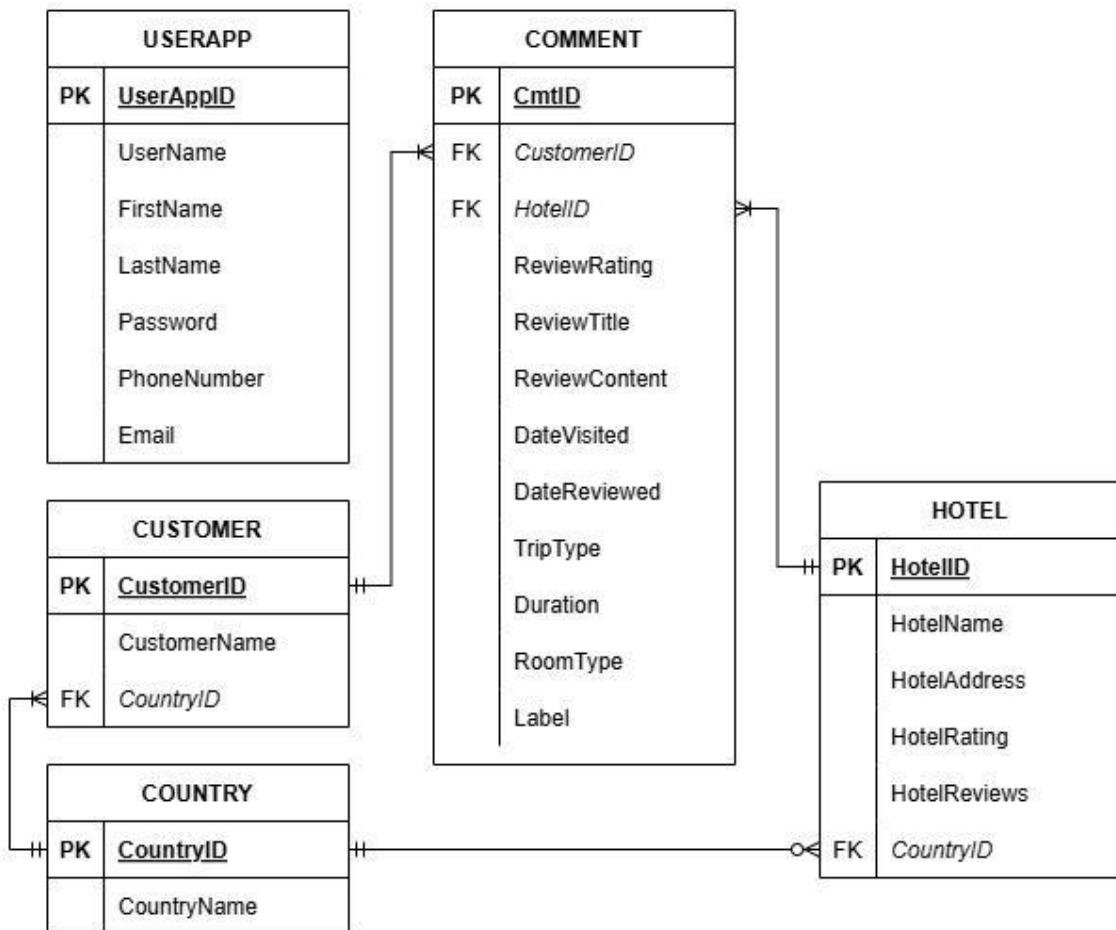
Sau khi ứng dụng vượt qua giai đoạn kiểm thử, nó sẽ được triển khai (Deployment). Ở bước này, ứng dụng được đưa lên môi trường thực tế, có thể là trên một máy chủ, một cửa hàng ứng dụng (App Store, Google Play), hoặc một nền tảng nội bộ dành cho doanh nghiệp. Đây là lúc người dùng thực sự có thể sử dụng ứng dụng.

Cuối cùng, quy trình này được thiết kế theo mô hình vòng lặp, có nghĩa là sau khi triển khai, ứng dụng sẽ tiếp tục được đánh giá để tìm ra những vấn đề mới và cải thiện dựa trên phản hồi của người dùng. Điều này có thể dẫn đến việc quay lại bước đầu tiên – xác định các điểm khó khăn mới để tiếp tục nâng cấp và tối ưu hóa ứng dụng.

## CHƯƠNG 4: MÔ HÌNH VÀ QUY TRÌNH THỰC HIỆN

### 4.1 Thiết kế cơ sở dữ liệu

#### 4.1.1 Entities and Categories



Hình 4.1: Database ERD

STT	Entity	Description
1	Userapp	Người dùng quản trị ứng dụng (Admin)
2	Customer	Khách hàng đã để lại đánh giá về khách sạn.
3	Hotel	Khách sạn được phân tích
4	Comment	Các bình luận và đánh giá mà khách hàng để lại cho khách sạn.
5	Country	Quốc gia mà khách sạn và khách hàng thuộc về.

Bảng 4.1: Các thực thể của Database

#### 4.1.1.1 Userapp

Attributes	Description	Categories
UserAppID	Admin's ID	Identifier
UserName	Tên đăng nhập	
FirstName	Tên của người dùng	
LastName	Họ của người dùng	
Password	Mật khẩu đăng nhập	
PhoneNumber	Số điện thoại Admin	
Email	Email Admin	

Bảng 4.2: Các thuộc tính của thực thể Userapp

#### 4.1.1.2 Customer

Attributes	Description	Categories
CustomerID	ID của khách hàng	Identifier
CustomerName	Tên khách hàng	
CountryID	Xác định quốc gia của khách hàng	

Bảng 4.3: Các thuộc tính của thực thể Customer

#### 4.1.1.3 Hotel

Attributes	Description	Categories
HotelID	ID của khách sạn	Identifier
HotelName	Tên khách sạn	
HotelAddress	Địa chỉ khách sạn	Composite
HotelRating	Đánh giá trung bình của khách sạn	
HotelReviews	Số lượng đánh giá của khách sạn.	
CountryID	Xác định quốc gia của	

	khách sạn	
--	-----------	--

Bảng 4.4: Các thuộc tính của thực thể Hotel

#### 4.1.1.4 Comment

Attributes	Description	Categories
CmtID	ID của bình luận	Identifier
CustomerID	Xác định khách hàng viết bình luận	
HotelID	Xác định khách sạn được đánh giá	
ReviewRating	Điểm đánh giá	
ReviewTitle	Tiêu đề của đánh giá	
ReviewContent	Nội dung chi tiết của đánh giá	
DateVisited	Ngày khách hàng ở khách sạn	Composite
DateReviewed	Ngày khách hàng viết đánh giá	
TripType	Loại chuyến đi	
RoomType	Loại phòng khách hàng ở	

Bảng 4.5: Các thuộc tính của thực thể Comment

#### 4.1.1.5 Country

Attributes	Description	Categories
CountryID	ID của quốc gia	
CountryName	Tên quốc gia	

Bảng 4.6: Các thuộc tính của thực thể Country

#### **4.1.2 Mô tả quy tắc nghiệp vụ**

##### **4.1.2.1 Userapp**

- Một người dùng ứng dụng có duy nhất một email và số điện thoại.
- Mỗi người dùng ứng dụng có một tài khoản với tên đăng nhập, mật khẩu để đăng nhập vào hệ thống.
- Người dùng ứng dụng có thể quản lý dữ liệu về khách sạn và đánh giá.

##### **4.1.2.2 Customer**

- Hai khách hàng không thể có cùng tên (CustomerName) và nước (CountryID).

##### **4.1.2.3 Hotel**

- Điểm đánh giá của khách sạn (HotelRating) được tính bằng trung bình cộng của tất cả các đánh giá (ReviewRating) từ khách hàng.
- Số lượng đánh giá của khách sạn (HotelReviews) là số lượng nhận xét đã được đăng tải.
- Hai khách sạn không thể có cùng tên (HotelName) và nước (CountryID).

##### **4.1.2.4 Comment**

- Một khách hàng có thể để lại nhiều đánh giá khác nhau cho cùng một khách sạn, nhưng mỗi đánh giá có một thời gian ghé thăm khác nhau (DateVisited).
- ReviewRating chỉ nhận các giá trị từ 1-10.
- Nếu khách sạn nào có comment thì số lượng comment tại thời điểm luôn tối thiểu là 50.

#### **4.1.3. Ràng buộc lực lượng của mô tả mối quan hệ**

Mối quan hệ	Loại	Mô tả
UserApp - Comment	1-n	Một người dùng ứng dụng có thể viết nhiều bình luận. Một bình luận chỉ thuộc về một người dùng.
Customer - Comment	1-n	Một khách hàng có thể viết nhiều bình luận. Một bình luận chỉ thuộc về một khách hàng.

Hotel - Comment	1-n	Một khách sạn có thể có nhiều bình luận. Một bình luận chỉ thuộc về một khách sạn.
Customer - Country	n-1	Một khách hàng đến từ một quốc gia. Một quốc gia có thể có nhiều khách hàng.
Hotel - Country	n-1	Một khách sạn nằm trong một quốc gia. Một quốc gia có thể có nhiều khách sạn.

Bảng 4.7: Mối quan hệ giữa các thực thể

#### 4.1.4. Thiết kế cơ sở dữ liệu vật lý

##### COUNTRY

Field	Data Types	Constraints
CountryID	INT	Primary Key, Auto Increment
CountryName	VARCHAR(100)	NOT NULL
CountryCode	VARCHAR(10)	NOT NULL, UNIQUE

Bảng 4.8: Cơ sở vật lý thực thể Country

##### CUSTOMER

Field	Data Types	Constraints
CustomerID	INT	Primary Key, Auto Increment
CustomerName	VARCHAR(100)	NOT NULL
CountryID	INT	Foreign Key References COUNTRY(CountryID) ON DELETE SET NULL

Bảng 4.9: Cơ sở vật lý thực thể Customer

## HOTEL

Field	Data Types	Constraints
HotelID	INT	Primary Key, Auto Increment
HotelName	VARCHAR(200)	NOT NULL
HotelAddress	VARCHAR(255)	NOT NULL
HotelRating	DECIMAL(3,1)	
HotelReviews	INT	DEFAULT 0
CountryID	INT	Foreign Key References COUNTRY(CountryID) ON DELETE CASCADE

Bảng 4.10: Cơ sở vật lý thực thể Hotel

## COMMENT

Field	Data Types	Constraints
CmtID	INT	Primary Key, Auto Increment
CustomerID	INT	Foreign Key References CUSTOMER(CustomerID) ON DELETE CASCADE
HotelID	INT	Foreign Key References HOTEL(HotelID) ON DELETE CASCADE
ReviewRating	DECIMAL(3,1)	CHECK (ReviewRating BETWEEN 1.0 AND 10.0)
ReviewTitle	VARCHAR(255)	

ReviewContent	TEXT	
DateVisited	DATE	
DateReviewed	DATE	
TripType	VARCHAR(50)	
RoomType	VARCHAR(50)	
Duration	TINYINT	

Bảng 4.11: Cơ sở vật lý thực thể Comment

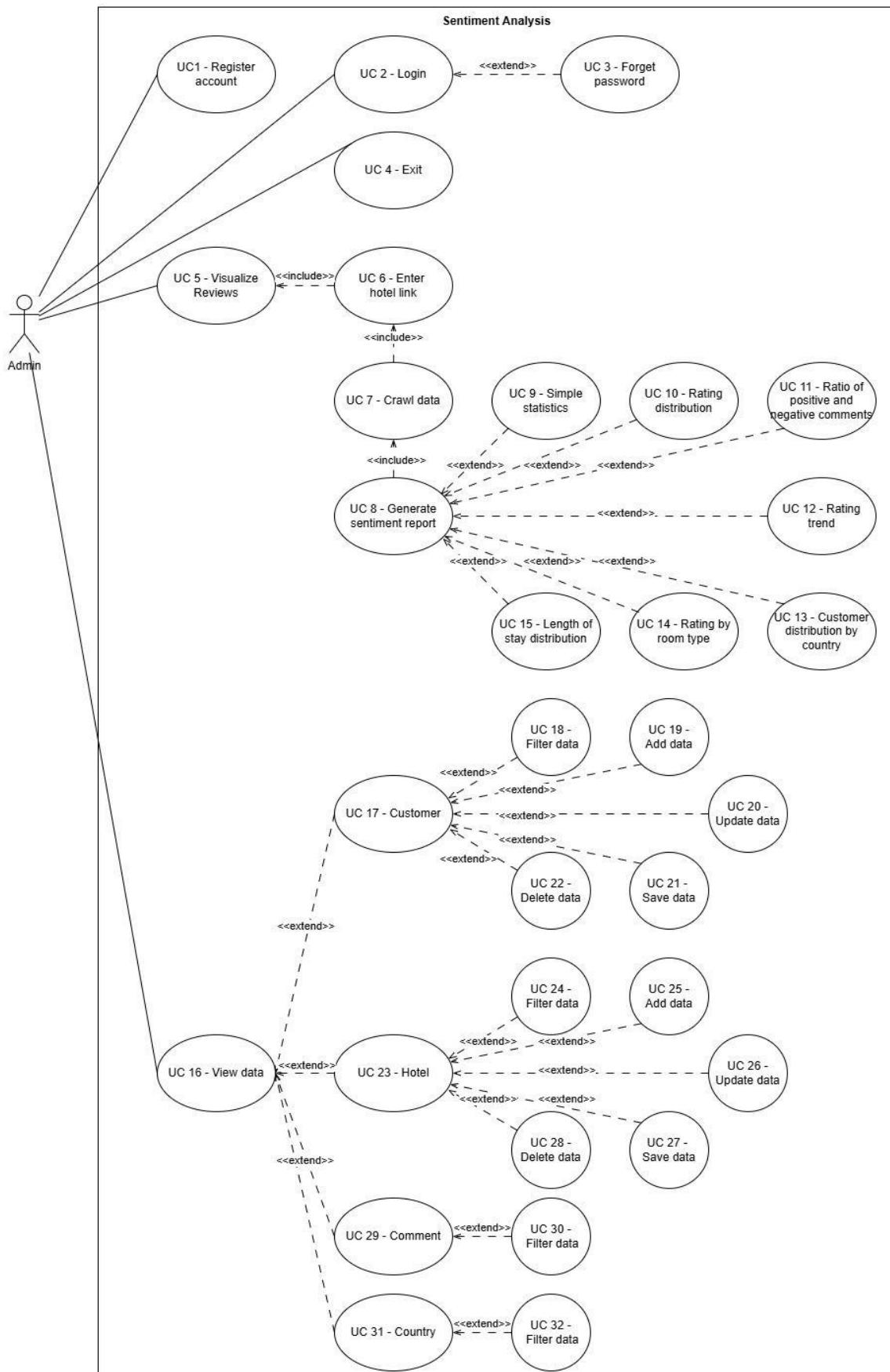
## 4.2 Sơ đồ Use Case

### 4.2.1 Sơ đồ Use Case

Ứng dụng Sentiment Analysis là một hệ thống giúp phân tích cảm xúc từ dữ liệu đánh giá khách sạn, sản phẩm hoặc dịch vụ. Hệ thống sử dụng các thuật toán xử lý ngôn ngữ tự nhiên (NLP) và học máy (Machine Learning) để phân tích nội dung đánh giá, trích xuất thông tin và phân loại cảm xúc thành các nhóm như tích cực, tiêu cực hoặc trung tính.

Người dùng chính của hệ thống là Admin, người có quyền đăng ký, đăng nhập, nhập liên kết khách sạn và xem báo cáo phân tích. Hệ thống tự động thu thập dữ liệu từ liên kết được cung cấp, thực hiện phân tích và hiển thị kết quả dưới dạng báo cáo trực quan. Các báo cáo này giúp người quản lý khách sạn hoặc doanh nghiệp hiểu rõ hơn về phản hồi của khách hàng, từ đó có thể cải thiện chất lượng dịch vụ.

Sơ đồ Use Case mô tả các chức năng chính của ứng dụng, bao gồm quá trình đăng ký, đăng nhập, nhập dữ liệu, tạo báo cáo và xem kết quả phân tích. Bên cạnh đó, hệ thống cũng hỗ trợ các tính năng mở rộng như khôi phục mật khẩu và xem chi tiết dữ liệu đánh giá.



Hình 4.2: Use case diagram

## 4.2.2 Mô tả Use Case

### 4.2.2.1 Tác nhân

Admin: Người dùng duy nhất trong hệ thống, có quyền truy cập và sử dụng các chức năng chính.

### 4.2.2.2 Các trường hợp sử dụng

UC 1 - Register account: Cho phép Admin đăng ký tài khoản.

UC 2 - Login: Cho phép Admin đăng nhập vào hệ thống.

- (Extend) UC 3 - Forget password: Cho phép đặt lại mật khẩu.

UC 4 - Exit: Thoát app.

UC 5 - Visualize Reviews: Trực quan hóa đánh giá.

- (Include) UC 6: Enter hotel link: Cho phép người dùng nhập URL của khách sạn cần phân tích.
- (Include) UC 7 - Crawl data: Thu thập dữ liệu từ link người dùng nhập.
- (Include) UC 8 - Generate sentiment report: Tạo báo cáo phân tích
  - (Extend) UC 9 - Simple statistics: Xem biểu đồ Thông kê đơn giản
  - (Extend) UC 10 - Rating distribution: Xem biểu đồ Phân bố điểm đánh giá.
  - (Extend) UC 11 - Ratio of positive and negative comments: Xem biểu đồ Tỉ lệ bình luận tích cực và tiêu cực.
  - (Extend) UC 12 - Rating trend: Xem biểu đồ Xu hướng điểm đánh giá
  - (Extend) UC 13 - Customer distribution by country: Xem biểu đồ Phân bố khách hàng theo quốc gia.
  - (Extend) UC 14 - Rating by room type: Xem biểu đồ Điểm đánh giá theo loại phòng.
  - (Extend) UC 15 - Length of stay distribution: Xem biểu đồ Phân phối thời gian lưu trú.

UC 16 - View data: Cho phép xem lại và điều chỉnh dữ liệu đã thu thập. Người dùng có thể xem, lọc và điều chỉnh dữ liệu của khách hàng và khách sạn; chỉ xem và lọc dữ liệu của bình luận và các nước.

- (Extend) UC 17 - Customer
  - (Extend) UC 18 - Filter data
  - (Extend) UC 19 - Add data
  - (Extend) UC 20 - Update data
  - (Extend) UC 21 - Save data
  - (Extend) UC 22 - Delete data
- (Extend) UC 23 - Hotel
  - (Extend) UC 18 - Filter data
  - (Extend) UC 19 - Add data
  - (Extend) UC 20 - Update data
  - (Extend) UC 21 - Save data
  - (Extend) UC 22 - Delete data
- (Extend) UC 23 - Comment
  - (Extend) UC 24 - Filter data
- (Extend) UC 25 - Country
  - (Extend) UC 26 - Filter data

## 4.3 Thiết kế giao diện người dùng

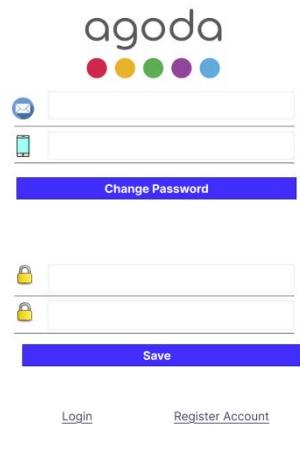
### 4.3.1 Giao diện đăng nhập

The screenshot shows the Agoda login page. At the top is the Agoda logo with five colored dots underneath. Below the logo are two input fields: one for email (with a user icon) and one for password (with a lock icon). To the right of the password field is a 'Show' checkbox. A large blue 'LOGIN' button is centered below the inputs. At the bottom of the form are links for 'Register Account' and 'Forget Password', and an 'Exit' link.

Hình 4.3: Giao diện đăng nhập

STT	Thành phần	Ý nghĩa
1	“LOGIN”	Sau khi nhập tài khoản bấm vào để đăng nhập và di chuyển tới giao diện người dùng.
2	“Show”	Hiển thị mật khẩu từ định dạng “***” đến định dạng chuỗi.
3	“Register Account”	Di chuyển đến giao diện đăng ký tài khoản khi chưa có tài khoản.
4	“Forget Password”	Di chuyển đến giao diện quên mật khẩu để lấy lại mật khẩu.

Bảng 4.12: Mô tả giao diện đăng nhập



*Hình 4.4: Giao diện quên mật khẩu*

STT	Thành phần	Ý nghĩa
1	“Change Password”	Xác nhận email và số điện thoại tài khoản trước để tiếp tục đổi mật khẩu.
2	“Save”	Lưu mật khẩu mới.
3	“Register Account”	Di chuyển đến giao diện đăng ký tài khoản khi chưa có tài khoản.
4	“Login”	Di chuyển đến giao diện đăng nhập.

*Bảng 4.13: Mô tả giao diện quên mật khẩu*

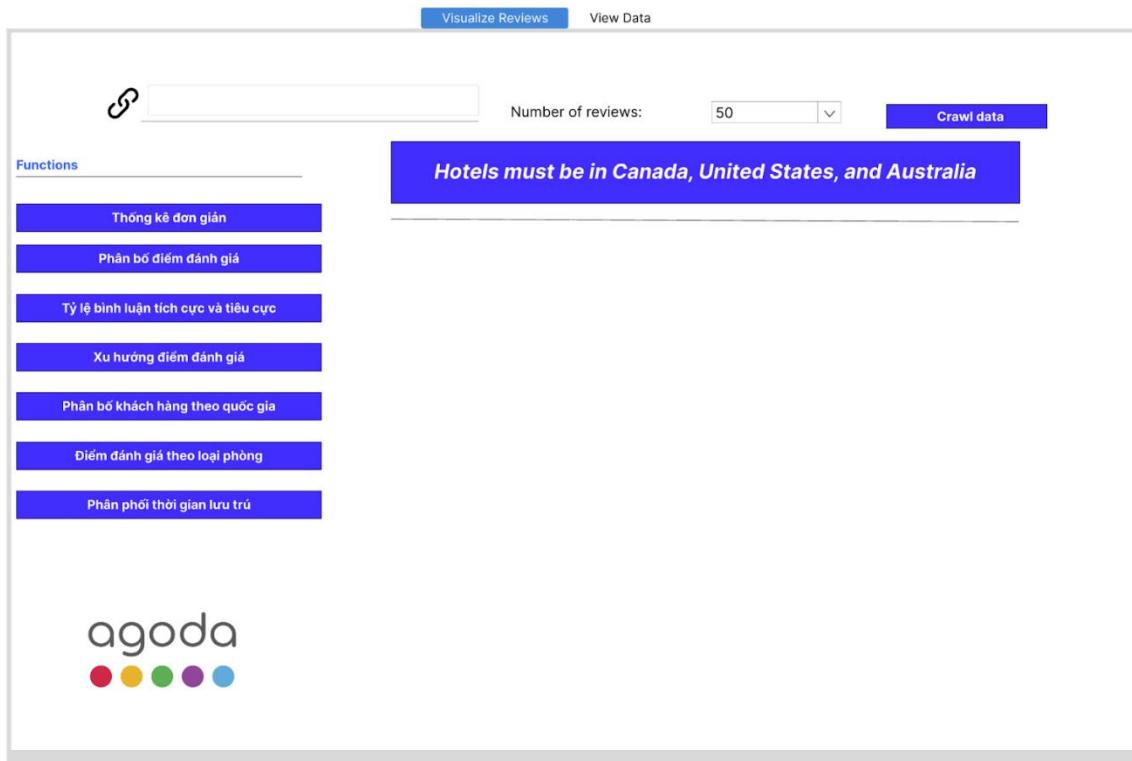


*Hình 4.5: Giao diện đăng ký tài khoản*

STT	Thành phần	Ý nghĩa
1	“Register”	Xác nhận đăng ký tài khoản mới.
2	“Forget Password”	Di chuyển đến giao diện quên mật khẩu.
4	“Login”	Di chuyển đến giao diện đăng nhập.

*Bảng 4.14: Mô tả giao diện đăng ký*

### 4.3.2 Giao diện sử dụng



Hình 4.6: Giao diện Visualize Reviews

STT	Thành phần	Ý nghĩa
1	“Crawl data”	Thực hiện chức năng crawl dữ liệu từ link nhập vào.
2	“Thống kê đơn giản”	Hiển thị các phân tích thống kê đơn giản từ dữ liệu.
3	“Phân bố điểm đánh giá”	Hiển thị biểu đồ phân bố điểm đánh giá.
4	“Tỷ lệ bình luận tích cực và tiêu cực”	Hiển thị biểu đồ tỷ lệ bình luận tích cực và tiêu cực
5	“Xu hướng điểm đánh giá”	Hiển thị biểu đồ xu hướng điểm đánh giá
6	“Phân bố khách hàng theo quốc gia”	Hiển thị biểu đồ phân bố khách hàng theo quốc gia

7	“Điểm đánh giá theo loại phòng”	Hiển thị biểu đồ điểm đánh giá theo loại phòng
8	“Phân phối thời gian lưu trú”	Hiển thị biểu đồ phân phối thời gian lưu trú

Bảng 4.15: Mô tả giao diện Visualize Reviews

Hình 4.7: Giao diện View Data Customers

Hình 4.8: Giao diện View Data Hotels

Hình 4.9: Giao diện View Data Countries

Hình 4.10: Giao diện View Data Comments

STT	Thành phần	Ý nghĩa
1	“Lọc”	Thực hiện lọc các dữ liệu theo ý muốn.
2	“Thêm”	Thực hiện thêm thông tin được nhập thêm.
3	“Lưu”	Thực hiện thay đổi thông tin sau khi chỉnh sửa.

4	“Cập nhật”	Thực hiện cập nhật danh sách sau khi chỉnh sửa hoặc thêm mới thông tin khách hàng.
5	“Xóa”	Thực hiện xóa thông tin đã chọn khỏi danh sách.

Bảng 4.16: Mô tả giao diện View Data

## CHƯƠNG 5: KẾT QUẢ THỰC NGHIỆM

### 5.1 Đánh giá mô hình Sentiment Analysis

Giai đoạn kiểm nghiệm cung cấp một đánh giá sơ lược về hiệu suất của từng thuật toán máy học. Các thuật toán này bao gồm Bernoulli NB, Random Forest Classifier, SVC và Logistic Regression.

Dựa trên bảng 5.1, có thể thấy rằng tất cả các bộ phân loại đều có độ chính xác tương đối khá cao, dao động từ 0.726783 đến 0.891552, trong đó Logistic Regression đạt độ chính xác cao nhất là 0.891552. Về độ chính xác (precision), LR tiếp tục đạt giá trị cao nhất là 0.887483, có nghĩa là các dự đoán của nó có tỉ lệ chính xác rất cao tuy vẫn có những trường hợp nằm ngoài dự báo đúng áy. Với Random Forest Classifier và SVC cũng có giá trị precision tương đối sát với LR, lần lượt là 0.854295 và 0.884722. Về chỉ số Recall, mô hình Logistic Regression tiếp tục đạt giá trị cao 0.886914, cho thấy khả năng phát hiện các trường hợp dương tính thực sự của nó là rất tốt. SVC cũng có Recall tương đương ở mức 0.884908, trong khi Random Forest Classifier đạt 0.835866, thấp hơn một chút nhưng vẫn duy trì hiệu suất ổn định. Xét về F1-Score, một chỉ số cân bằng giữa Precision và Recall, Logistic Regression vẫn dẫn đầu với 0.887195, theo sát là SVC với 0.884814 và Random Forest Classifier với 0.842503. Mô hình Bernoulli NB có hiệu suất thấp hơn đáng kể so với các mô hình còn lại, với Accuracy, Precision, Recall và F1-Score đều ở mức thấp nhất.

	Model	Accuracy	Precision	Recall	F1 Score
0	Bernoulli NB	0.726783	0.739270	0.746089	0.726200
1	Random Forest Classifier	0.852007	0.854295	0.835866	0.842503
2	SVC	0.889155	0.884722	0.884908	0.884814
3	Logistic Regression	0.891552	0.887483	0.886914	0.887195

Bảng 5.1: So sánh các thuật toán

Từ bảng trên, ta nhận thấy rằng Logistic Regression nổi bật với chỉ số Accuracy, Precision, Recall và F1-Score đều dao động cao trong khoảng mức 0.88 đến 0.89. Điều này chứng tỏ rằng LR là mô hình hoạt động hiệu quả nhất trong bộ dữ liệu này, với khả năng cân bằng tốt ở bốn chỉ số. So sánh với các mô hình khác, SVC và

Random Forest Classifier cũng có hiệu suất tương đương với LR, chỉ chênh lệch nhẹ nhưng vẫn duy trì một mức độ ổn định và hiệu suất đáng tin cậy. Tóm lại, Logistic Regression được coi là lựa chọn tối ưu nhất nhờ vào hiệu suất toàn diện, trong khi SVC và Random Forest có thể là những phương án thay thế phù hợp.

Giai đoạn kiểm nghiệm cung cấp một đánh giá chi tiết về hiệu suất của hai thuật toán máy học: Logistic Regression và LSTM. Hai mô hình này được so sánh dựa trên bốn tiêu chí chính: Accuracy, Precision, Recall và F1-Score. Dữ liệu trong bảng 5.2 cho thấy mỗi mô hình có những điểm mạnh riêng, nhưng LSTM nổi bật hơn trong bối cảnh phân tích cảm xúc người dùng qua bình luận.

	Model	Accuracy	Precision	Recall	F1 Score
0	Logistic Regression	0.891552	0.887483	0.886914	0.887195
1	LSTM	0.900539	0.909785	0.847578	0.877581

Bảng 5.2: So sánh Logistic Regression và LSTM

Dựa vào bảng 5.2, có thể thấy rằng LSTM đạt độ chính xác (Accuracy) cao hơn so với Logistic Regression, với giá trị 0.900539 so với 0.891552. Điều này cho thấy mô hình LSTM có khả năng phân loại chính xác hơn một chút so với Logistic Regression. Một độ chính xác cao hơn đồng nghĩa với việc giảm thiểu các dự đoán sai, giúp cải thiện hiệu suất tổng thể của hệ thống.

Xét về độ chính xác (Precision), LSTM đạt 0.909785, cao hơn đáng kể so với 0.887483 của Logistic Regression. Precision đo lường mức độ chính xác của các dự đoán dương tính, cho thấy rằng trong số các dự đoán được gán nhãn tích cực, LSTM có tỷ lệ dự đoán đúng cao hơn. Điều này rất quan trọng trong việc phân tích cảm xúc, đặc biệt khi cần đảm bảo rằng các bình luận được dự đoán là tích cực thực sự mang tính chất đó.

Tuy nhiên, xét về chỉ số Recall, Logistic Regression lại có lợi thế với 0.886914 so với 0.847578 của LSTM. Điều này có nghĩa là Logistic Regression có khả năng phát hiện các trường hợp dương tính thực sự cao hơn. Trong bối cảnh phân tích cảm xúc, Recall rất quan trọng khi cần xác định đầy đủ các bình luận tích cực hoặc tiêu cực.

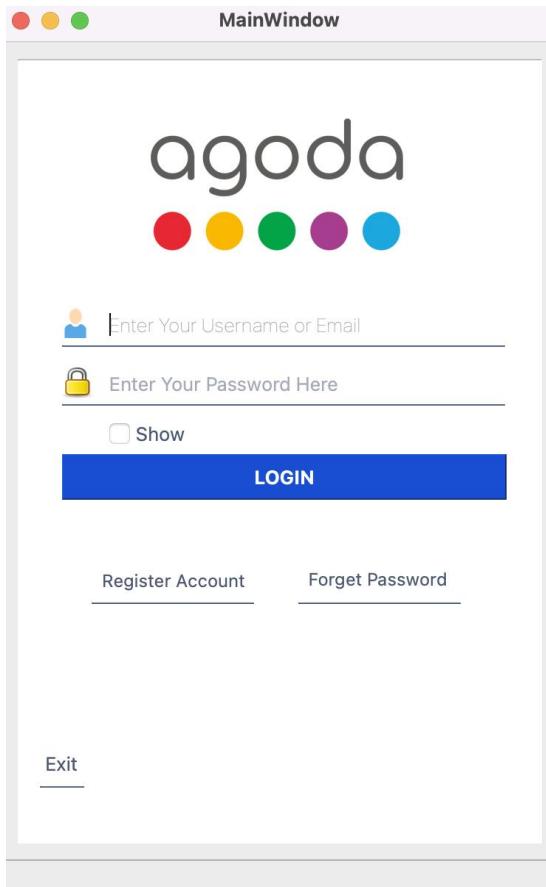
Việc LSTM có chỉ số Recall thấp hơn có thể chỉ ra rằng một số bình luận thuộc nhóm dương tính thực sự đã bị bỏ sót.

Về chỉ số F1-Score, một chỉ số cân bằng giữa Precision và Recall, Logistic Regression đạt 0.887195 trong khi LSTM có giá trị 0.877581. Mặc dù Logistic Regression có F1-Score nhỉnh hơn một chút, nhưng với sự vượt trội của LSTM về Accuracy và Precision, sự chênh lệch này không ảnh hưởng đáng kể đến việc lựa chọn mô hình tối ưu.

Tóm lại, mặc dù Logistic Regression có chỉ số Recall cao hơn, nhưng xét về hiệu suất tổng thể, LSTM vượt trội hơn nhờ độ chính xác cao nhất và Precision vượt trội. Đặc biệt, trong bài toán phân tích cảm xúc từ bình luận, mô hình cần có khả năng xác định chính xác các bình luận có cảm xúc tích cực hoặc tiêu cực mà không bị nhiễu bởi những dự đoán sai. Vì vậy, nhóm chúng tôi quyết định lựa chọn LSTM là mô hình tối ưu cho bài toán phân tích cảm xúc người dùng qua bình luận, nhờ vào khả năng học hỏi sâu sắc từ dữ liệu văn bản và hiệu suất tổng thể xuất sắc.

## 5.2 Trải nghiệm người dùng

### 5.2.1 Màn hình đăng nhập



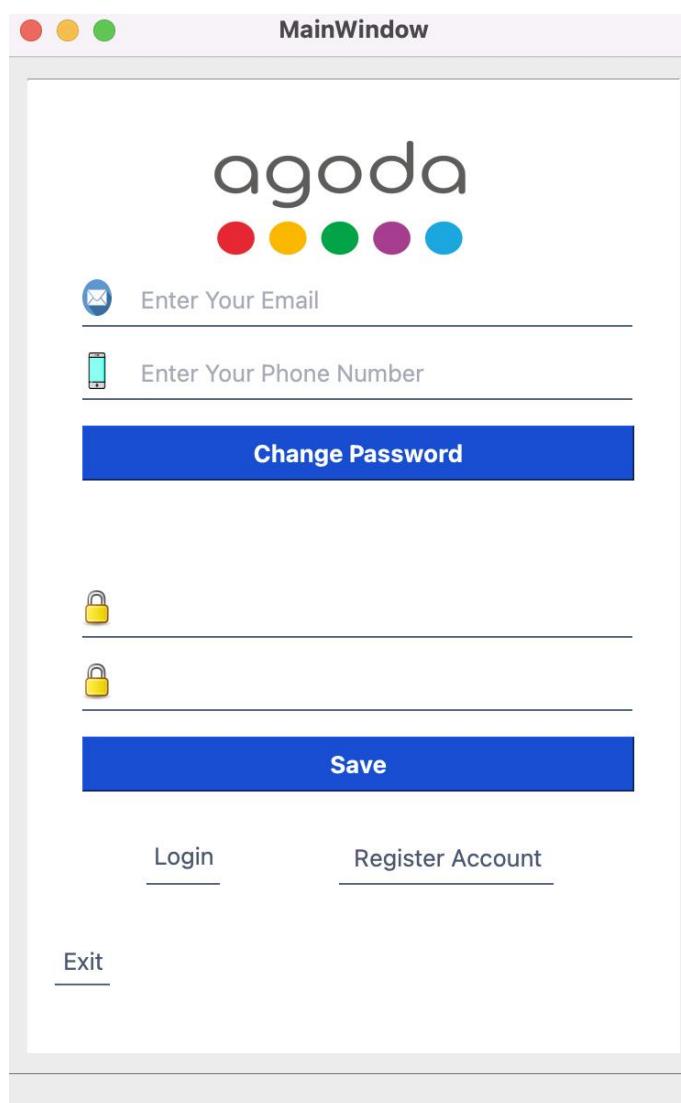
Hình 5.1: Màn hình đăng nhập

*Mô tả:* Người dùng bắt đầu từ trang đăng nhập, nhập tên đăng nhập và mật khẩu để truy cập vào hệ thống.

*Hành động:* Nhập tên đăng nhập và mật khẩu, sau đó nhấn nút "Đăng nhập". Nếu thiếu một trong hai thông tin, hệ thống sẽ hiển thị thông báo: "**Bạn cần điền đầy đủ thông tin**", yêu cầu người dùng nhập đầy đủ dữ liệu. Nếu thông tin đăng nhập không chính xác, thông báo "**Thông tin không chính xác**" sẽ xuất hiện và người dùng phải nhập lại.

*Chuyển tiếp:* Khi đăng nhập thành công, người dùng sẽ được chuyển hướng đến **Trang chủ**.

### 5.2.2 Quên mật khẩu



Hình 5.2: Màn hình quên mật khẩu

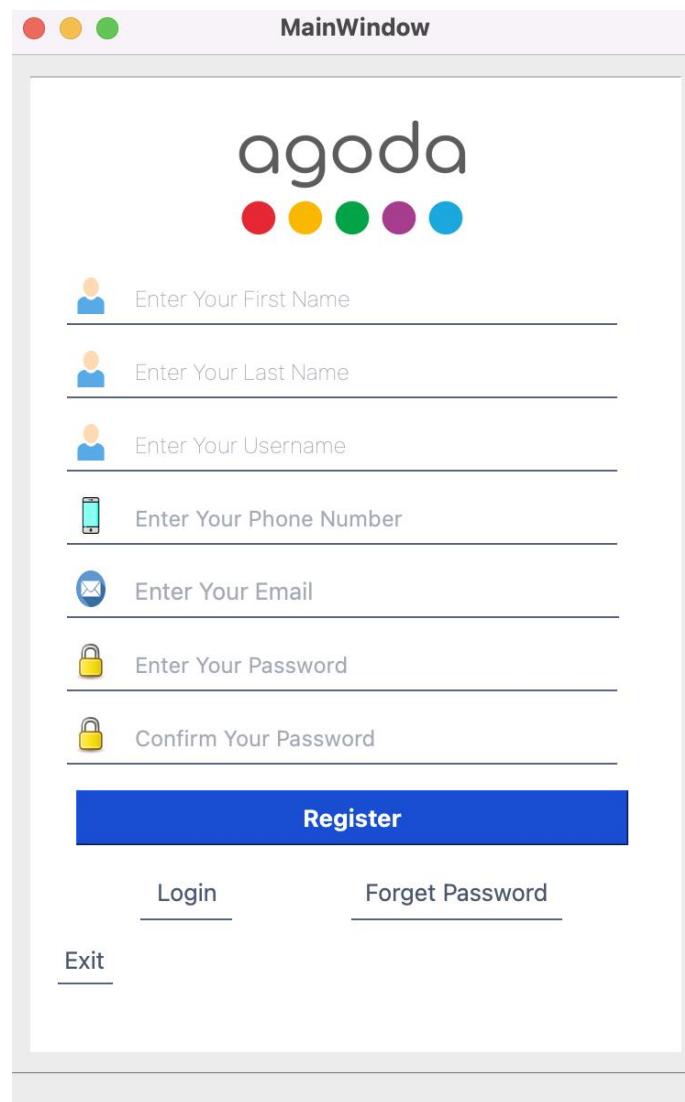
*Mô tả:* Người dùng bắt đầu tại trang **Quên mật khẩu**, nhập Email và Số điện thoại trùng khớp với thông tin tài khoản hiện có. Nếu người dùng nhập sai thì hệ thống không chuyển hướng đến phần đổi mật khẩu. Nếu người dùng nhập đúng, hệ thống chuyển hướng người dùng đến phần **Đổi mật khẩu**, nơi họ nhập mật khẩu mới. Khi quá trình đặt lại mật khẩu hoàn tất, người dùng sẽ được chuyển về trang **Đăng nhập** để truy cập vào hệ thống.

*Hành động của người dùng:* Nhập **Email** và **Số điện thoại** trùng khớp với tài khoản hiện có sau đó nhập **Mật khẩu mới** và **xác nhận mật khẩu mới**, nhấn **Lưu** để đặt lại mật khẩu.

- Nếu bất kỳ trường **Email hoặc Số điện thoại** bị bỏ trống, hệ thống sẽ hiển thị thông báo: "**Bạn cần điền đầy đủ thông tin**", yêu cầu người dùng nhập đủ dữ liệu.
- Nếu **Mật khẩu mới** và **Xác nhận mật khẩu** không trùng khớp, thông báo "**Mật khẩu xác nhận không khớp với mật khẩu đã nhập**" sẽ xuất hiện, yêu cầu người dùng nhập lại mật khẩu chính xác.

*Chuyển tiếp:* Khi đặt lại mật khẩu thành công, người dùng sẽ được chuyển đến trang **Đăng nhập**, sau đó đăng nhập vào **Trang chủ** của hệ thống.

### 5.2.3 Đăng ký tài khoản



Hình 5.3: Màn hình đăng ký tài khoản

*Mô tả:* Người dùng bắt đầu tại trang **Đăng ký**, nhập **Họ, Tên, Tên đăng nhập, Số điện thoại, Email, Mật khẩu** và xác nhận mật khẩu để tạo tài khoản. Sau khi tài

khoản được tạo thành công, hệ thống sẽ tự động chuyển hướng người dùng đến trang **Đăng nhập** để truy cập vào hệ thống.

**Hành động:** Nhập **Họ, Tên, Tên đăng nhập, Số điện thoại, Email** và **Mật khẩu**, sau đó nhấn nút "**Đăng ký**" để hoàn tất quá trình tạo tài khoản.

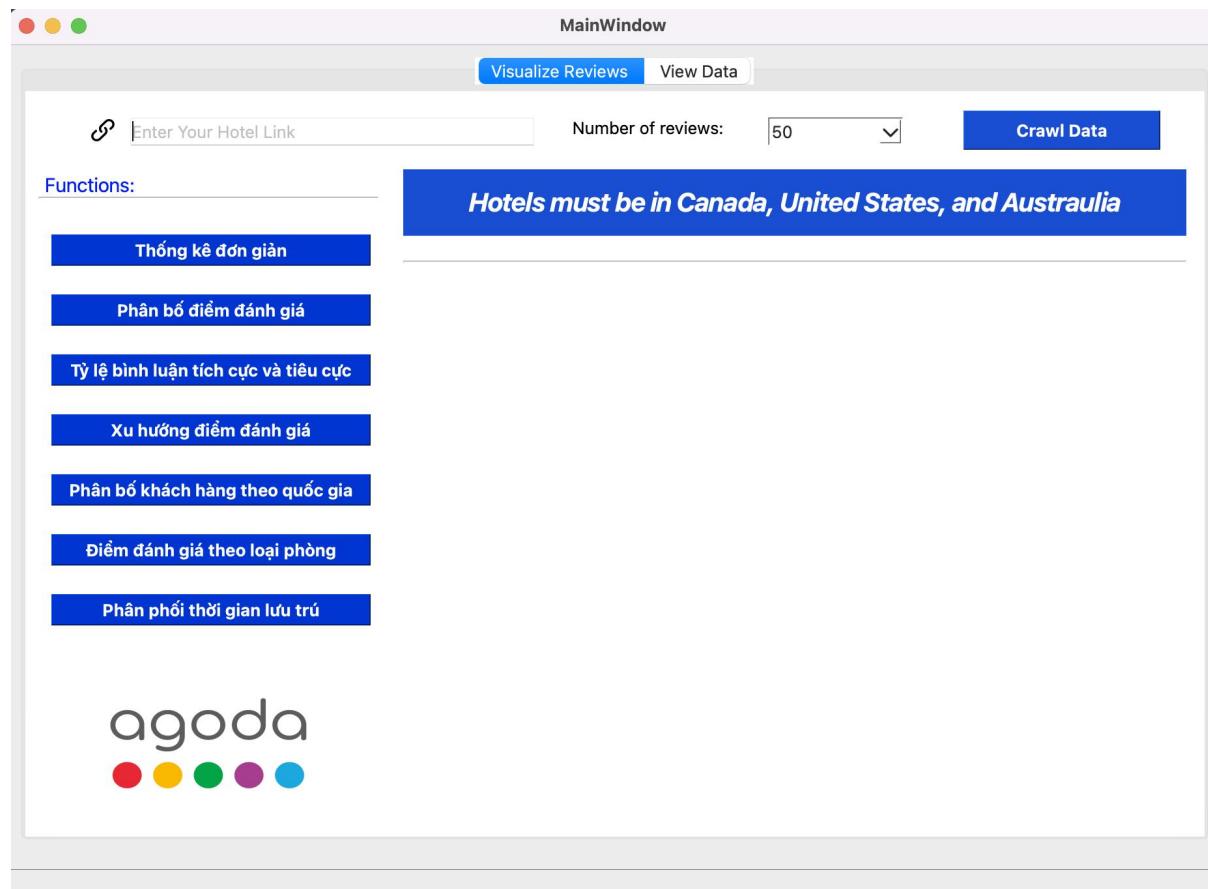
- Nếu bất kỳ trường nào trong **Họ, Tên** hoặc **Tên đăng nhập** bị bỏ trống, hệ thống sẽ hiển thị thông báo: "**Bạn cần điền đầy đủ thông tin**", yêu cầu người dùng nhập đầy đủ dữ liệu.

- Nếu **Mật khẩu mới** và **Xác nhận mật khẩu** không trùng khớp, thông báo "**Mật khẩu xác nhận không khớp với mật khẩu đã nhập**" sẽ xuất hiện, yêu cầu người dùng nhập lại chính xác.

**Chuyển tiếp:** Sau khi đăng ký thành công, người dùng sẽ được chuyển đến trang **Đăng nhập**, sau đó có thể đăng nhập vào **Trang chủ** của hệ thống.

#### 5.2.4 Trang chủ

##### 5.2.4.1 Trực quan hóa đánh giá



Hình 5.4: Trang chủ

*Mô tả:*

- Người dùng có thể dán link khách sạn vào ô nhập để thu thập dữ liệu.
- Lựa chọn số lượng đánh giá: Người dùng có thể chọn số lượng đánh giá muốn crawl từ một danh sách thả xuống.
- Nút "Crawl Data": Kích hoạt quá trình thu thập dữ liệu từ link khách sạn đã nhập.
  - Thông báo giới hạn địa điểm: Ứng dụng chỉ hỗ trợ khách sạn ở Canada, Hoa Kỳ và Úc.

*Các chức năng chính:*

- Thống kê đơn giản - Hiển thị các chỉ số cơ bản về đánh giá.
- Phân bố điểm đánh giá - Trực quan hóa mức điểm của các đánh giá.
- Tỷ lệ bình luận tích cực và tiêu cực – So sánh tỷ lệ đánh giá tốt/xấu.
- Xu hướng điểm đánh giá – Phân tích sự thay đổi của điểm số theo thời gian.
- Phân bố khách hàng theo quốc gia – Xác định nguồn gốc của khách hàng.
- Điểm đánh giá theo loại phòng – So sánh đánh giá giữa các loại phòng khác nhau.
- Phân phối thời gian lưu trú – Phân tích thời gian khách thường lưu trú.

*Ví dụ với khách sạn Orleans Hotel and Casino*

Home > United States Hotels (388,007) > Las Vegas (NV) Hotels (1,113) > Las Vegas (NV) Resorts (86) > Book Orleans Hotel and Casino [See all 1,113 properties in Las Vegas \(NV\)](#)

**Orleans Hotel and Casino** ★★★★  
4500 W. Tropicana Ave, West of The Strip, Las Vegas (NV), United States, 89103 - [SEE MAP](#)

Experience the vibrant energy of Las Vegas' iconic West of The Strip neighborhood at Orleans Hotel and Casino. Indulge in luxury with a refreshing swim in the private pool, try your luck at the exclusive casino, and pamper yourself at the on-site salon. Unwind in comfortable rooms with modern amenities, complimentary Wi-Fi, and captivating city views. Located in a clean and secure area, the friendly staff will ensure you have a memorable stay. Plus, enjoy the best bargain off the Strip. Perfect for two travelers seeking excitement and relaxation in the heart of Las Vegas. [Some content may be Generative AI assisted. Inaccuracies may occur.]

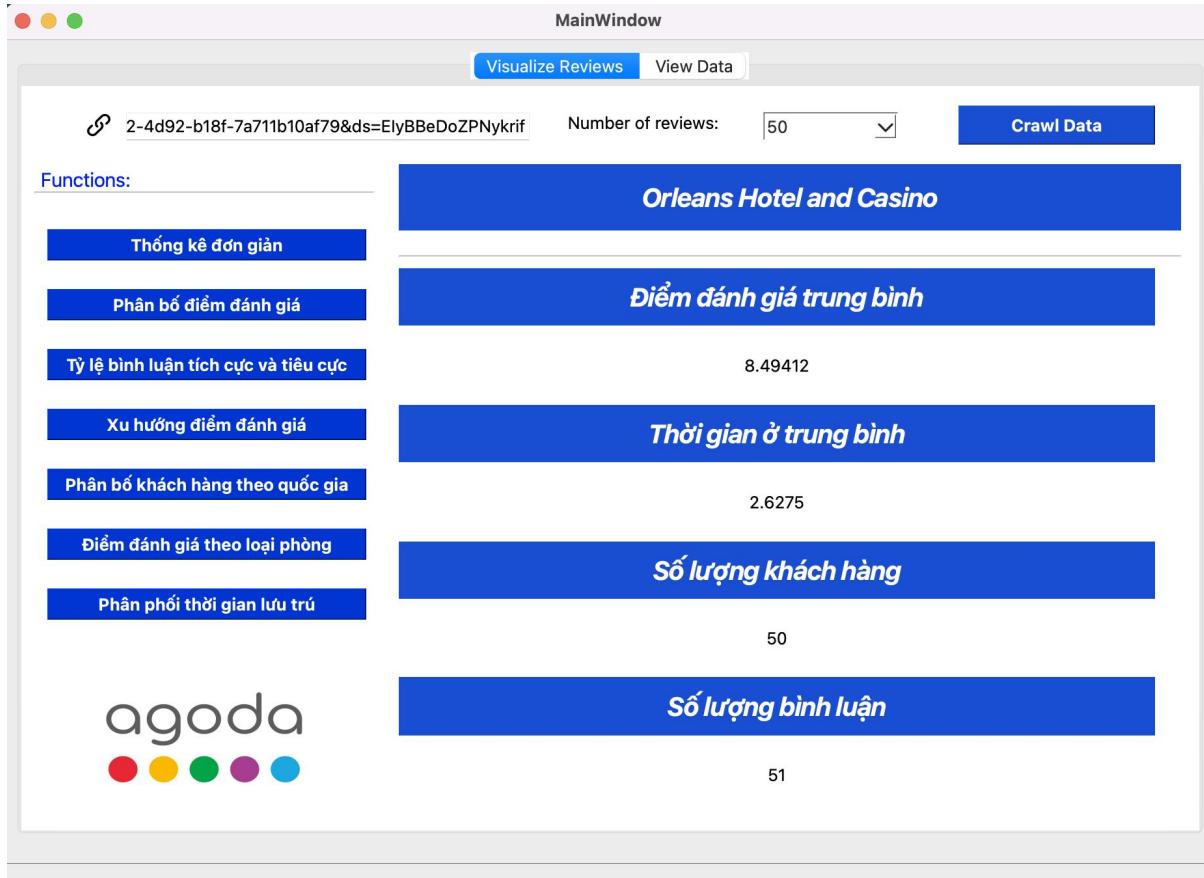
**8.1 Excellent** [Read all reviews](#)  
907 reviews

Room comfort and quality 9.0 Service 8.4  
Cleanliness 8.3 Value for money 8.3 ⓘ

"great stay" "Be >"

Hình 5.5: Khách sạn Orleans Hotel and Casino

- Thống kê đơn giản



Hình 5.6: Minh họa về thống kê đơn giản

Dữ liệu thống kê đơn giản cho khách sạn **Orleans Hotel and Casino** cung cấp cái nhìn tổng quan về chất lượng dịch vụ và hành vi của khách hàng. Cụ thể, khách sạn này đạt **điểm đánh giá trung bình là 8.49412**, phản ánh mức độ hài lòng cao từ khách lưu trú. **Thời gian lưu trú trung bình của khách hàng là 2.6275 đêm**, cho thấy phần lớn du khách có xu hướng nghỉ lại trong khoảng thời gian ngắn. Tổng số khách hàng được ghi nhận là **50 người**, với **51 bình luận** phản hồi về trải nghiệm tại khách sạn. Những số liệu này không chỉ giúp đánh giá mức độ phổ biến của khách sạn mà còn hỗ trợ trong việc phân tích xu hướng và cải thiện chất lượng dịch vụ.

#### - Phân bố điểm đánh giá

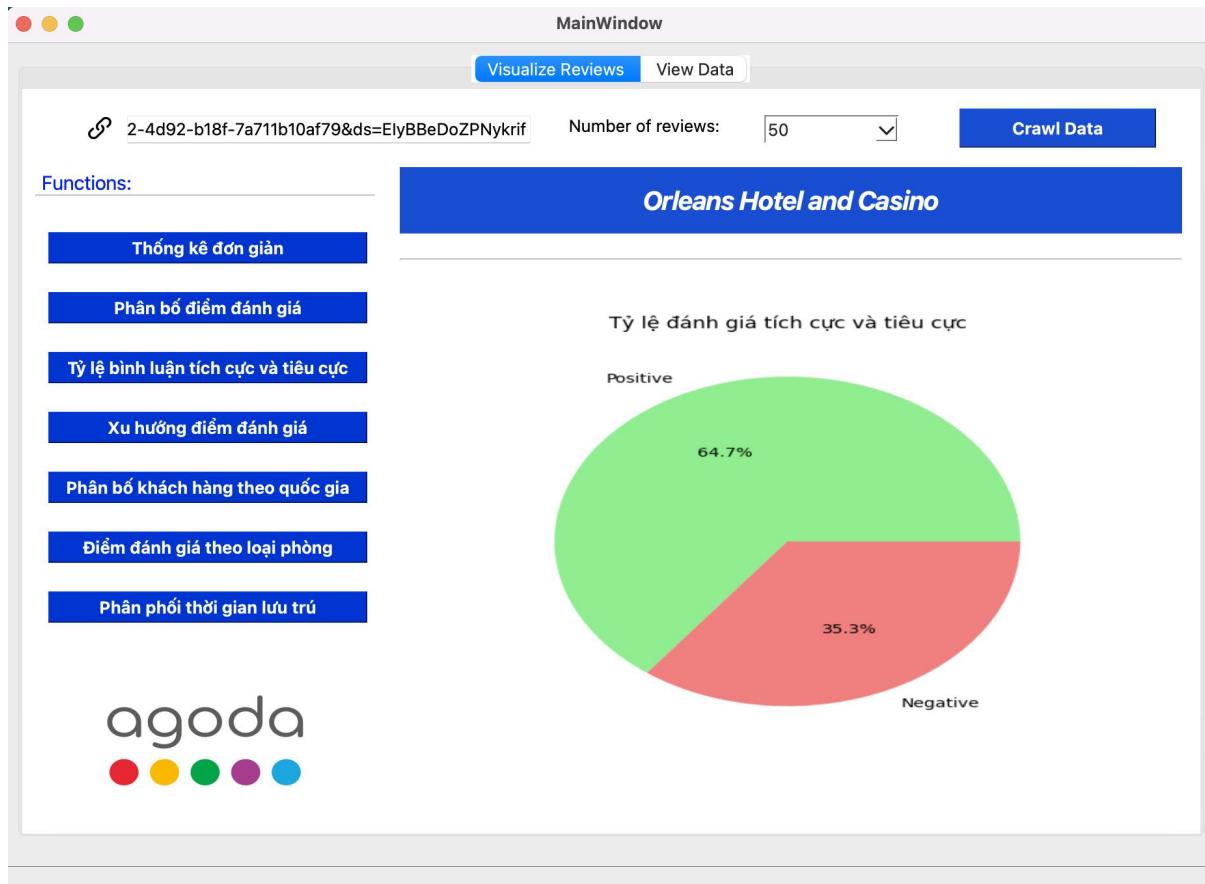


Hình 5.7: Minh họa về phân bố điểm đánh giá

Biểu đồ **Phân bố điểm đánh giá** cho khách sạn **Orleans Hotel and Casino** cho thấy đa số đánh giá tập trung vào mức điểm cao. Cụ thể, phần lớn khách hàng chấm điểm trong khoảng từ **8 đến 10**, với số lượng đáng kể nhất ở mức **9 và 10**, thể hiện sự hài lòng cao về chất lượng dịch vụ. Trong khi đó, các mức điểm thấp hơn từ **1 đến 6** xuất hiện rất ít, cho thấy chỉ có một số ít khách hàng không hài lòng với trải nghiệm tại khách sạn. Đường cong mật độ trên biểu đồ cũng phản ánh xu hướng phân bố lệch

về phía điểm cao, cung cấp nhận định rằng Orleans Hotel and Casino nhận được đánh giá tích cực từ khách hàng. Những thông tin này có thể giúp khách sạn duy trì chất lượng dịch vụ và tiếp tục cải thiện để thu hút nhiều khách hơn trong tương lai.

- Tỷ lệ bình luận tích cực tiêu cực



Hình 5.8: Minh họa về tỷ lệ đánh giá tích cực và tiêu cực

Biểu đồ **Tỷ lệ đánh giá tích cực và tiêu cực** của khách sạn **Orleans Hotel and Casino** cho thấy phần lớn phản hồi từ khách hàng là tích cực. Cụ thể, **64.7%** số đánh giá là **tích cực**, thể hiện sự hài lòng cao của khách hàng đối với dịch vụ tại khách sạn. Trong khi đó, **35.3%** số đánh giá là **tiêu cực**, phản ánh một số vấn đề còn tồn tại cần được cải thiện.

Tỷ lệ đánh giá này cho thấy Orleans Hotel and Casino có chất lượng dịch vụ tốt với đa số khách hàng có trải nghiệm tích cực. Tuy nhiên, tỷ lệ phản hồi tiêu cực vẫn chiếm hơn một phần ba tổng số bình luận, điều này gợi ý rằng khách sạn cần phân tích sâu hơn các đánh giá tiêu cực để xác định nguyên nhân và cải thiện dịch vụ, nâng cao trải nghiệm khách hàng trong tương lai.

- Xu hướng điểm đánh giá

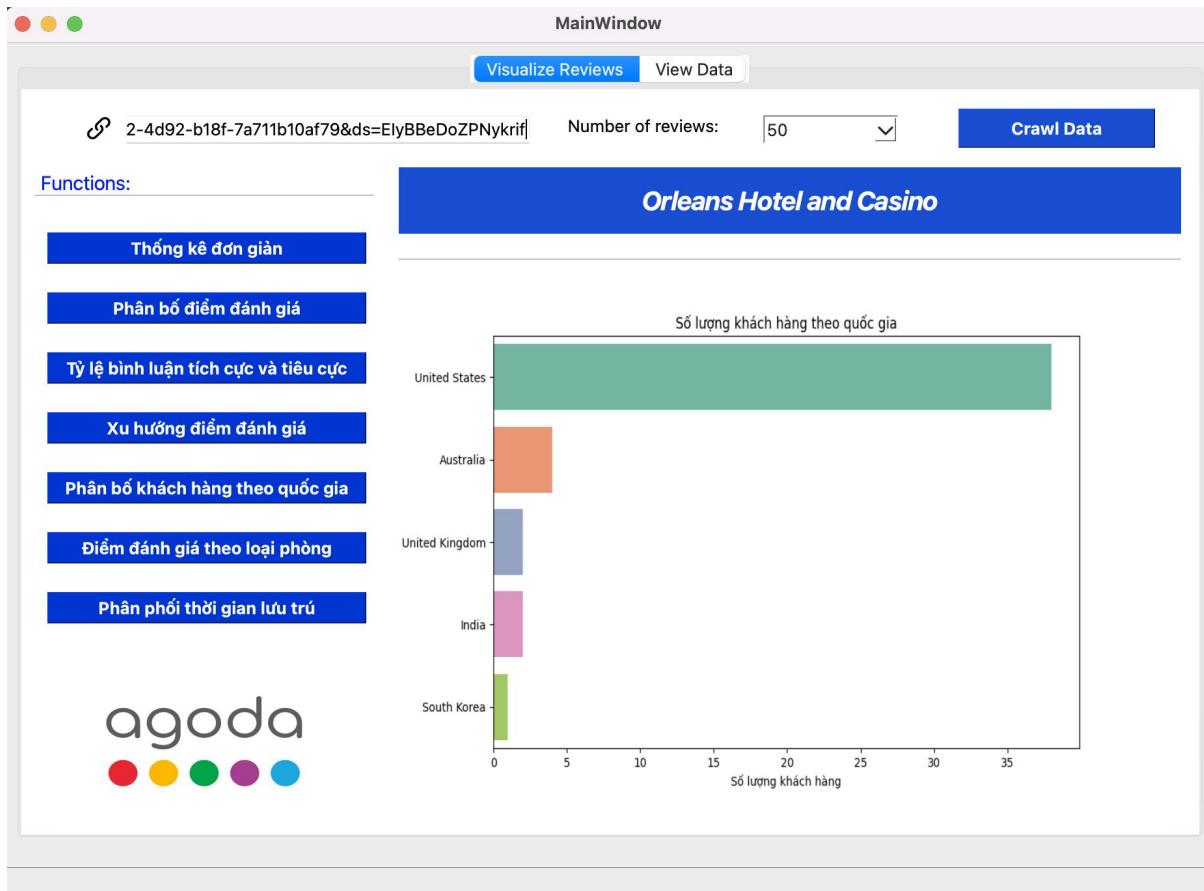


Hình 5.9: Minh họa về xu hướng điểm đánh giá

Biểu đồ **Xu hướng điểm đánh giá theo thời gian** của khách sạn **Orleans Hotel and Casino** cho thấy sự biến động đáng kể trong mức độ hài lòng của khách hàng. Mặc dù có những thời điểm đánh giá đạt mức cao gần **10**, nhưng cũng có một số giai đoạn điểm số giảm mạnh xuống mức thấp nhất.

Xu hướng này cho thấy sự không ổn định trong trải nghiệm của khách hàng tại khách sạn. Nguyên nhân có thể đến từ sự thay đổi trong chất lượng dịch vụ, các yếu tố mùa vụ hoặc những sự kiện đặc biệt ảnh hưởng đến sự hài lòng của khách hàng. Để cải thiện tình hình, Orleans Hotel and Casino có thể xem xét kỹ lưỡng các giai đoạn có điểm số giảm mạnh, phân tích nguyên nhân và đưa ra các biện pháp khắc phục nhằm nâng cao sự ổn định và chất lượng dịch vụ trong dài hạn.

- Phân bố khách hàng theo quốc gia

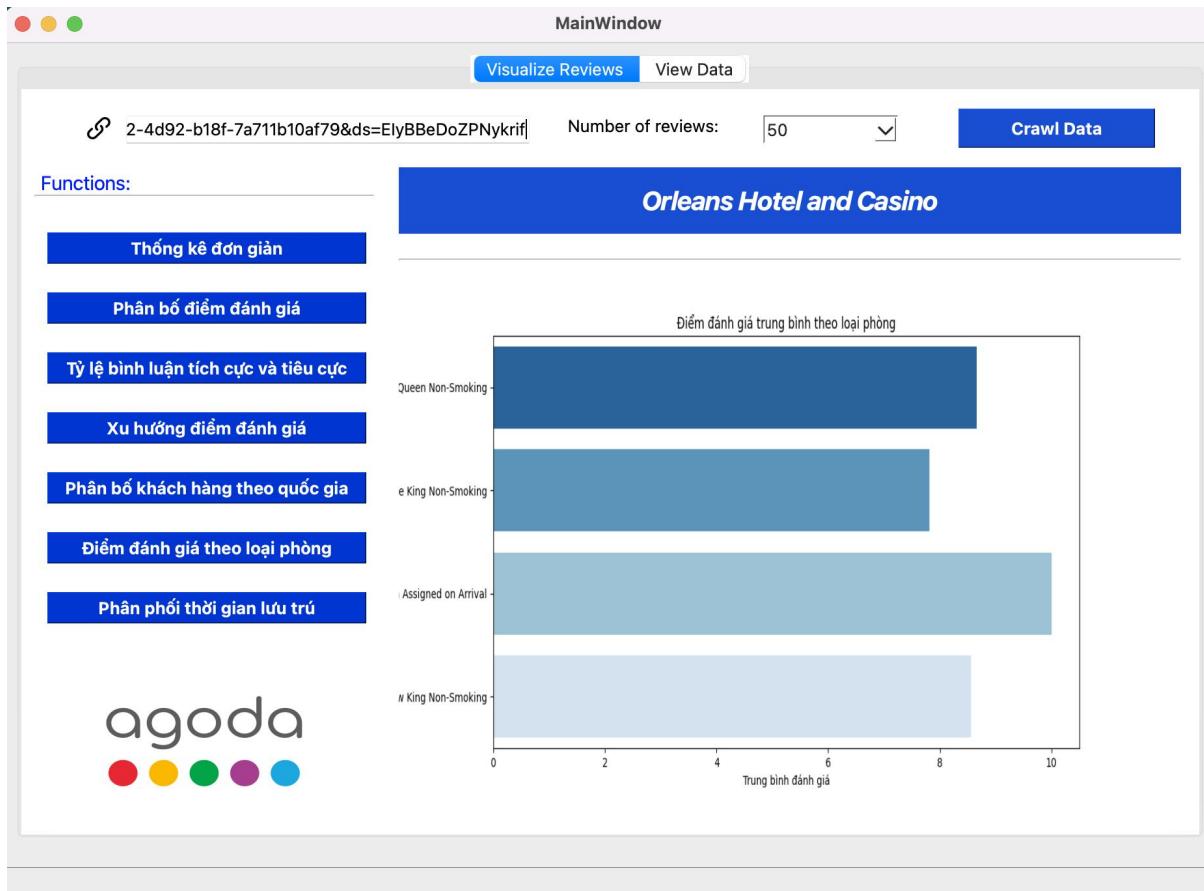


Hình 5.10: Minh họa về phân bố khách hàng theo quốc gia

Biểu đồ **Phân bố khách hàng theo quốc gia** của khách sạn **Orleans Hotel and Casino** cho thấy phần lớn du khách đến từ **Hoa Kỳ**, chiếm số lượng áp đảo so với các quốc gia khác. Điều này có thể phản ánh sự phổ biến của khách sạn đối với khách nội địa hoặc vị trí địa lý thuận lợi thu hút du khách Mỹ.

Ngoài ra, một số lượng nhỏ khách đến từ **Úc, Vương quốc Anh, Ấn Độ và Hàn Quốc**, cho thấy khách sạn cũng thu hút một nhóm khách quốc tế nhưng chưa thực sự đa dạng. Để mở rộng thị trường và thu hút thêm du khách nước ngoài, khách sạn có thể xem xét triển khai các chiến lược quảng bá phù hợp với từng khu vực, cải thiện dịch vụ đa ngôn ngữ hoặc hợp tác với các nền tảng đặt phòng phổ biến ở những quốc gia này.

- Phân bố theo loại phòng

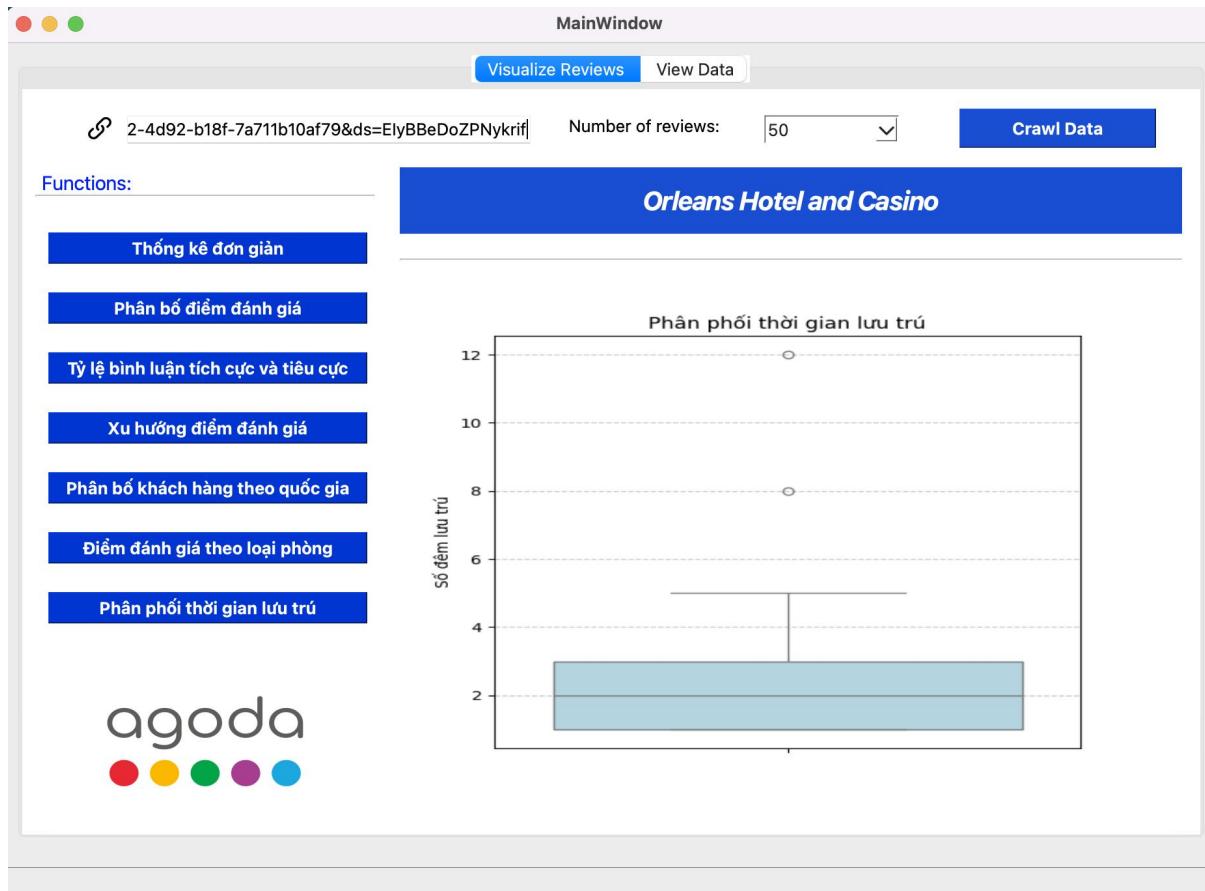


Hình 5.11: Minh họa về phân bố theo loại phòng

Biểu đồ **Điểm đánh giá trung bình theo loại phòng** của khách sạn **Orleans Hotel and Casino** cho thấy sự phân bố điểm đánh giá giữa các loại phòng khác nhau. Các phòng **Queen Non Smoking** và **King Non Smoking** nhận được điểm đánh giá trung bình cao nhất, lần lượt đạt mức 9,3 và 9,0, cho thấy khách hàng đánh giá khá cao về sự thoải mái và chất lượng dịch vụ ở các loại phòng này.

Trong khi đó, các phòng **Assigned on Arrival** có điểm đánh giá trung bình thấp hơn, khoảng 8,1, phản ánh sự khác biệt trong trải nghiệm của khách khi nhận phòng. Mặc dù vậy, các loại phòng này vẫn duy trì mức đánh giá tốt, chứng tỏ chất lượng dịch vụ của khách sạn nhìn chung là ổn định. Thông qua kết quả này, khách sạn có thể cân nhắc cải thiện thêm các loại phòng có điểm đánh giá thấp để nâng cao trải nghiệm của khách hàng.

- Phân phối thời gian lưu trú



Hình 5.12: Minh họa về phân phối thời gian lưu trú

Biểu đồ **Phân phối thời gian lưu trú** của khách sạn **Orleans Hotel and Casino** thể hiện sự phân bổ số đêm khách lưu trú tại khách sạn. Phần lớn khách hàng có thời gian lưu trú dao động trong khoảng từ **1 đến 4 đêm**, với giá trị trung vị khoảng **2,6 đêm**. Điều này cho thấy đa số khách đến khách sạn có xu hướng lưu trú ngắn hạn, có thể là khách du lịch hoặc khách đi công tác ngắn ngày.

Ngoài ra, biểu đồ cũng cho thấy một số giá trị ngoại lai, với một số ít khách lưu trú hơn **10 đêm**. Đây có thể là những trường hợp khách ở dài ngày vì mục đích nghỉ dưỡng hoặc công tác dài hạn. Thông tin này có thể giúp khách sạn tối ưu hóa chiến lược giá và dịch vụ để phục vụ tốt hơn nhu cầu của từng nhóm khách hàng khác nhau.

#### 5.2.4.2 Hiển thị dữ liệu

Chức năng **hiển thị dữ liệu** trong giao diện của bạn được thiết kế để giúp người dùng xem và quản lý thông tin một cách dễ dàng.

	Customer Name	Country
1	Juanita	United States
2	Mae	United States
3	Michael	United States
4	Isaac	United States
5	Erin	United States
6	Jaycee	United States
7	Jan	United States
8	Steven	United States
9	PRINCESS-...	United States
10	Nelson	United States
11	Dennis	United States
12	Robert	United States
13	Kenneth	United States
14	Maxwell	United States
15	Jennifer	United States

Hình 5.13: Hiển thị dữ liệu khách hàng

Mô tả:

Các tab lựa chọn loại dữ liệu, bao gồm **Customers**, **Hotels**, **Countries**, **Comments**. Tab **Customers** đang được mở, hiển thị danh sách khách hàng theo quốc gia. Bên trái là bộ lọc dữ liệu cho phép người dùng chọn quốc gia từ một danh sách thả xuống. Hiện tại, quốc gia được chọn là **United States**.

Bên dưới bộ lọc là khu vực nhập thông tin khách hàng với hai ô nhập liệu: một để nhập hoặc chỉnh sửa tên khách hàng, một để nhập hoặc chỉnh sửa quốc gia của họ. Các nút chức năng đi kèm giúp người dùng thực hiện các thao tác quản lý dữ liệu khách hàng.

Các chức năng chính:

- Lọc dữ liệu: Người dùng có thể chọn quốc gia để tìm kiếm dữ liệu trong danh sách.
- Thêm dữ liệu: Cho phép nhập thông tin khách hàng mới và thêm vào danh sách.
- Lưu thông tin: Lưu các thay đổi khi chỉnh sửa thông tin khách hàng.

- Cập nhật dữ liệu: Cập nhật danh sách sau khi chỉnh sửa hoặc thêm mới thông tin khách hàng.

- Xóa dữ liệu: Xóa thông tin khách hàng đã chọn khỏi danh sách.

The screenshot shows a software application window titled "MainWindow". At the top, there are tabs for "Customers", "Hotels" (which is selected), "Countries", and "Comments". Below the tabs, there is a search interface on the left with the following fields:

- Hotel Country: A dropdown menu showing "Australia" with a "▼" icon.
- Hotel Name: An empty text input field.
- Hotel Country: An empty text input field.
- Hotel Address: An empty text input field.
- Hotel Rating: A numeric input field with a value of "0,00" and up/down arrows.
- Hotel Reviews: A numeric input field with a value of "0" and up/down arrows.

Below these fields are four buttons arranged in a grid:

Thêm	Lưu
Cập nhật	Xoá

On the right side of the window, there is a table titled "Hotels" with the following data:

	Hotel Name	Hotel Country	Hotel Address	Hotel Rating
1	Rydges ...	Australia	8 Arrivals ...	8.6
2	Amora ...	Australia	11 Jamison ...	8.8

Hình 5.14: Hiển thị dữ liệu khách sạn

#### Khu vực bộ lọc và nhập liệu (bên trái):

Người dùng có thể lọc khách sạn theo quốc gia bằng cách chọn từ danh sách thả xuống (hiện tại là "Australia") và nhấn nút "**Lọc**". Điều này giúp tìm kiếm nhanh chóng các khách sạn thuộc một quốc gia cụ thể.

Biểu mẫu nhập liệu bao gồm: Hotel Name, Hotel Country, Hotel Address, Hotel Rating và Hotel Reviews. Dưới biểu mẫu nhập liệu là bốn nút chức năng chính: thêm, lưu, cập nhật và xóa khách sạn.

Khu vực bảng hiển thị danh sách khách sạn (bên phải): Bảng này gồm bốn cột chính: **Hotel Name** hiển thị tên khách sạn, **Hotel Country** thể hiện quốc gia, **Hotel Address** ghi địa chỉ cụ thể, và **Hotel Rating** cung cấp điểm đánh giá. Dữ liệu trong

bảng được hiển thị theo từng dòng, với các khách sạn như "Rydges" và "Amora" ở Australia, có địa chỉ cụ thể và điểm đánh giá tương ứng là 8.6 và 8.8.

The screenshot shows a software application window titled "MainWindow". At the top, there are three buttons: "Visualize Reviews", "View Data" (which is highlighted in blue), and "Customers", "Hotels", "Countries", "Comments". Below these buttons is a search/filtering section with a green header labeled "Lọc". It contains two input fields: "Country Name:" and "Country Code:". To the right of this is a large table displaying a list of countries with their names and corresponding ISO codes. The table has columns for "CountryName" and "CountryCode". The data is sorted by CountryName. The first 15 rows of the table are as follows:

	CountryName	CountryCode
1	Afghanistan	AF
2	Åland Islands	AX
3	Albania	AL
4	Algeria	DZ
5	American ...	AS
6	Andorra	AD
7	Angola	AO
8	Anguilla	AI
9	Antarctica	AQ
10	Antigua & ...	AG
11	Argentina	AR
12	Armenia	AM
13	Aruba	AW
14	Australia	AU
15	Austria	AT

Hình 5.15: Hiển thị danh sách quốc gia

*Khu vực lọc và nhập liệu (bên trái):* Đây là nơi người dùng có thể tìm kiếm thông tin quốc gia theo tên hoặc mã quốc gia. Khu vực này chứa một nút "Lọc", giúp lọc dữ liệu theo tiêu chí nhập vào. Dưới nút lọc là hai trường nhập liệu: "**Country Name**" để nhập hoặc tìm kiếm theo tên quốc gia, và "**Country Code**" để tìm kiếm theo mã quốc gia hai chữ cái.

*Bảng hiển thị danh sách quốc gia (bên phải):* Bảng này có hai cột chính: "**Country Name**" hiển thị tên quốc gia, và "**Country Code**" thể hiện mã quốc gia theo chuẩn ISO 3166-1 alpha-2. Bảng dữ liệu được sắp xếp theo thứ tự bảng chữ cái của tên quốc gia, với các quốc gia như Afghanistan (AF), Albania (AL), Algeria (DZ), Australia (AU), và Austria (AT). Cột mã quốc gia có biểu tượng mũi tên, cho thấy dữ liệu có thể được sắp xếp theo thứ tự tăng dần hoặc giảm dần.

MainWindow

Visualize Reviews   **View Data**

**Customers**   **Hotels**   **Countries**   **Comments**

Date Visited	01/12/2023	<input type="button" value="From"/>	01/12/2024	<input type="button" value="To"/>
Trip Type	Couple			
	<input type="button" value="Lọc"/>			
Customer Name:				
Hotel Name:				
Review Rating:				
Review Title:				
Date Visited:				
Date Reviewed:				
Room Type:				
Duration:				
Label				

	Customer Name	Hotel Name	Review Rating	Review Title
1	Juanita	Orleans Hote...	8.0	Only for on
2	Mae	Orleans Hote...	7.6	Good locat
3	Erin	Orleans Hote...	8.4	Quick ...
4	Nelson	Orleans Hote...	8.4	Great for ...
5	Kenneth	Orleans Hote...	5.2	Buyer bewi
6	Maxwell	Orleans Hote...	8.0	Very good
7	Jennifer	Orleans Hote...	10.0	Great "olde
8	Helen	Orleans Hote...	10.0	Exceptiona
9	Jesse	Orleans Hote...	8.0	Very good
10	나윤	Orleans Hote...	8.8	Good
11	RODERICK	Orleans Hote...	10.0	Best Time
12	Georgette	Orleans Hote...	10.0	Exceptiona
13	Frank	Orleans Hote...	10.0	Sr. Fun Tim
14	Aniket	Orleans Hote...	10.0	Good
15	Remy	Rosen Inn at ...	8.0	Pardon the v

*Hình 5.16: Hiển thị danh sách các bình luận đánh giá của khách hàng về các khách sạn*

**Khu vực lọc và nhập liệu (bên trái):** Đây là nơi người dùng có thể nhập thông tin để tìm kiếm và lọc dữ liệu đánh giá. Một bộ lọc thời gian cho phép người dùng chọn phạm vi "**Date Visited**", với hai trường ngày bắt đầu và ngày kết thúc. Ngoài ra, người dùng có thể lọc dữ liệu theo loại hình chuyến đi (**Trip Type**). Khi người dùng nhập thông tin và nhấn nút "**Lọc**", bảng dữ liệu bên phải sẽ cập nhật để hiển thị các đánh giá phù hợp với tiêu chí lọc. Bên dưới bộ lọc là biểu mẫu nhập liệu, bao gồm các trường để nhập hoặc chỉnh sửa thông tin đánh giá. Các trường này bao gồm **Customer Name** để nhập tên khách hàng, **Hotel Name** để ghi tên khách sạn được đánh giá, **Review Rating** để nhập điểm số đánh giá, **Review Title** để ghi tiêu đề đánh giá, **Date Visited** để nhập ngày khách hàng lưu trú, **Date Reviewed** để nhập ngày đánh giá được đăng, **Room Type** để ghi loại phòng, **Duration** để nhập thời gian lưu trú, và **Label** để thêm nhãn hoặc thông tin phân loại bổ sung.

**Bảng hiển thị danh sách đánh giá (bên phải):** Bảng này có nhiều cột chứa thông tin chi tiết về các đánh giá, bao gồm **Customer Name** (tên khách hàng), **Hotel Name** (tên khách sạn được đánh giá), **Review Rating** (điểm đánh giá trên thang điểm 10), và

**Review Title** (tiêu đề hoặc nội dung ngắn gọn của đánh giá). Các dữ liệu hiển thị cho thấy khách sạn "Orleans Hotel" có nhiều đánh giá với các điểm số khác nhau, từ 5.2 đến 10, và nhiều nhận xét ngắn gọn như "Only for on", "Great locat", "Best Time", "Good", và "Pardon the...". Một số khách sạn khác cũng xuất hiện trong danh sách, cho thấy hệ thống có thể lưu trữ và hiển thị đánh giá từ nhiều khách sạn khác nhau.

## CHƯƠNG 6: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN TRONG TƯƠNG LAI

### 6.1 Kết luận

Trong dự án này, nhóm đã phát triển một ứng dụng phân tích cảm xúc tự động dựa trên các kỹ thuật Xử lý ngôn ngữ tự nhiên (NLP) và Học máy (ML), nhằm hỗ trợ các doanh nghiệp trong ngành khách sạn hiểu rõ hơn về phản hồi của khách hàng. Ứng dụng được thiết kế để thu thập, xử lý và phân tích dữ liệu đánh giá từ nền tảng trực tuyến, cung cấp thông tin về mức độ hài lòng và xu hướng của khách hàng.

Các mô hình học máy truyền thống như Naïve Bayes, Random Forest, Logistic Regression và Support Vector Classifier đã được triển khai để so sánh hiệu suất, cùng với mô hình học sâu LSTM. Kết quả thực nghiệm cho thấy, mô hình LSTM có khả năng nhận diện cảm xúc tốt hơn so với các mô hình truyền thống, nhờ vào khả năng xử lý ngữ cảnh của chuỗi dữ liệu.

Bên cạnh đó, hệ thống cũng được tích hợp với cơ sở dữ liệu để lưu trữ thông tin và cung cấp giao diện giúp người dùng dễ dàng khai thác dữ liệu. Việc áp dụng kỹ thuật trực quan hóa dữ liệu giúp nâng cao hiệu quả phân tích và hỗ trợ ra quyết định chiến lược cho doanh nghiệp.

### 6.2 Hạn chế và hướng phát triển trong tương lai

Mặc dù ứng dụng đã đạt được nhiều kết quả tích cực, nhưng vẫn tồn tại một số hạn chế cần được khắc phục. Thứ nhất, ứng dụng chưa thể phân tích sâu các bình luận có ngữ nghĩa phức tạp, ví dụ hoặc mang tính mỉa mai, điều này có thể dẫn đến sai lệch trong việc phân loại cảm xúc của người dùng. Để cải thiện, có thể tích hợp các mô hình ngôn ngữ tiên tiến như BERT hoặc GPT nhằm xử lý tốt hơn các bình luận có tính chất phức tạp và bối cảnh sâu.

Thứ hai, ứng dụng hiện mới chỉ hỗ trợ phân tích cảm xúc từ các bình luận bằng tiếng Anh, chưa hỗ trợ đa ngôn ngữ, gây hạn chế trong việc mở rộng thị trường. Hướng phát triển trong tương lai là nâng cấp hệ thống để có thể phân tích cảm xúc trên nhiều ngôn ngữ khác nhau, giúp tiếp cận đối tượng khách hàng đa dạng hơn.

Cuối cùng, một số mô hình học máy vẫn chưa được tối ưu hóa hoàn toàn về mặt tốc độ và tài nguyên tính toán, đặc biệt khi làm việc với lượng dữ liệu lớn, điều này có thể ảnh hưởng đến thời gian xử lý và khả năng mở rộng của hệ thống. Để khắc phục, có thể ứng dụng các kỹ thuật tối ưu hóa mô hình như giảm kích thước mô hình, áp

dụng kỹ thuật song song hóa hoặc sử dụng phần cứng chuyên dụng như GPU để cải thiện tốc độ xử lý.

## TÀI LIỆU THAM KHẢO

- [1] Amazon Web Services (AWS). (n.d.). Xử lí ngôn ngữ tự nhiên (NLP) là gì? Truy cập từ <https://aws.amazon.com/vi/what-is/nlp/>
- [2] Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. International Conference on Learning Representations (ICLR).
- [3] Berkson, J. (1944). Application of the Logistic Function to Bio-Assay. Journal of the American Statistical Association.
- [4] Boyd, S., & Vandenberghe, L. (2004). Convex Optimization. Cambridge University Press.
- [5] CodeGym. (n.d.). Database là gì? 8 mô hình Database phổ biến hiện nay. Truy cập từ <https://codegym.vn/blog/database-la-gi/>
- [6] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL.
- [7] Elmasri, R., & Navathe, S. B. (2020). Fundamentals of Database Systems (7th ed.). Pearson.
- [8] Gaurav, M., Krishna Kumar, M., & Sunil, K. (2023, April). Sentiment analysis on images using convolutional neural networks based Inception-V3 transfer learning approach.
- [9] GeeksforGeeks. (n.d.). Bernoulli naive Bayes.  
<https://www.geeksforgeeks.org/bernoulli-naive-bayes/>
- [10] GeeksforGeeks. (n.d.). Random forest algorithm in machine learning. Retrieved from <https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/>
- [11] Graves, A., Mohamed, A. R., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing.
- [12] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning (2nd ed.). Springer.
- [13] He, X., Shi, B., Bai, X., Xia, G. S., Zhang, Z., & Dong, W. (2019). Image caption generation with part of speech guidance. Pattern Recognition Letters, 119, 229–237.

- [14] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780.
- [15] Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *Proceedings of ACL*.
- [16] Jurafsky, D., & Martin, J. H. (2021). *Speech and Language Processing* (3rd ed.). Pearson.
- [17] Kaspersky. (n.d.). Trợ lý ảo Alexa, Siri và Google Assistant có sử dụng AI không?<https://kaspersky.proguide.vn/san-pham-cong-nghe-moi/tro-ly-ao-alexa-siri-va-google-assistant-co-su-dung-ai-khong/>
- [18] Lê, H. (n.d.). Rừng ngẫu nhiên (Random Forest). *Machine Learning Cơ Bản*. [https://machinelearningcoban.com/tabml\\_book/ch\\_model/random\\_forest.html](https://machinelearningcoban.com/tabml_book/ch_model/random_forest.html)
- [19] Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.
- [20] Pabel, A., & Prideaux, B. (2016). Social media use in pre-trip planning by tourists visiting a small regional leisure destination. *Journal of Vacation Marketing*, 22(4), 335-348.
- [21] Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*.
- [22] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*.
- [23] Punithavathi, R., Manoharan, P., Garima, S., & Kumar, C. (2024, May). Transforming sentiment analysis for e-commerce product reviews: Hybrid deep learning model with an innovative term weighting and feature selection.
- [24] Qin, Y., Song, D., Chen, H., Cheng, W., Jiang, G., & Cottrell, G. (2017). A dual-stage attention-based recurrent neural network for time series prediction. *Advances in Neural Information Processing Systems*.
- [25] Scikit-learn. (n.d.). Naive Bayes classifiers. Truy cập từ [https://scikit-learn.org/stable/modules/naive\\_bayes.html](https://scikit-learn.org/stable/modules/naive_bayes.html)
- [26] TheTelegraph. (2013). Tripadvisor and the issue of trust.
- [27] TopDev. (n.d.). Database là gì? Các kiểu Database phổ biến và ứng dụng. <https://topdev.vn/blog/database-la-gi-cac-kiem-database-pho-bien-va-ung-dung/>

[28] Vietnix. (n.d.). Database là gì? Các mô hình Database phổ biến và ứng dụng hiện nay. Truy cập từ <https://vietnix.vn/database-la-gi/>

[29] 200Lab. (n.d.). Natural Language Processing (NLP) là gì và ứng dụng của NLP. <https://200lab.io/blog/natural-language-processing-nlp-la-gi-va-ung-dung-cua-nlp>

[30] Wikipedia. (n.d.). *Sigmoid function* [Image]. Truy cập từ <https://de.m.wikipedia.org/wiki/Datei:Sigmoid-function-2.svg>

[31] GeeksforGeeks. (n.d.). *Random Forest* [Image]. Truy cập từ <https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/>