



agoda



ĐỀ TÀI

Sentiment analysis app

GVHD: Th.S Nguyễn Quang Phúc

THÀNH VIÊN NHÓM

K234060688

Lê Hữu Đăng

K224101336

Lê Nguyễn Minh Tài

K234060727

Phạm Quốc Thắng

K234060739

Bùi Thị Thu Vân

K234060723

Phan Vũ Khánh Quỳnh

OVERVIEW

- 01
- 02
- 03
- 04
- 05
- 06

TỔNG QUAN ĐỀ TÀI

CƠ SỞ LÝ THUYẾT

PHƯƠNG PHÁP ĐỀ XUẤT

MÔ HÌNH VÀ QUY TRÌNH THỰC HIỆN

KẾT QUẢ THỰC NGHIỆM

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN





TỔNG QUAN





Động lực nghiên cứu



Phản hồi trực tuyến ảnh hưởng lớn đến quyết định đặt phòng, đòi hỏi khách sạn hiểu rõ tâm lý khách hàng.



Ứng dụng NLP giúp tự động phân tích cảm xúc, tối ưu hóa dịch vụ và chiến lược kinh doanh, tạo lợi thế cạnh tranh bền vững.





Câu hỏi nghiên cứu



Làm sao áp dụng hiệu quả NLP và ML để phân tích cảm xúc từ bình luận sản phẩm?



Làm sao thiết kế module crawl dữ liệu thống nhất, đảm bảo độ chính xác và liên tục?



Làm sao tích hợp cơ sở dữ liệu với hệ thống phân tích để hỗ trợ chiến lược kinh doanh?



Đối tượng & phạm vi

Đối tượng

Các bình luận đánh giá trên website **agoda**
<https://www.agoda.com/>



Phạm vi thời gian

Không giới hạn khoảng thời gian, từ dữ liệu cũ
cho đến mới nhất.

Phạm vi không gian

Bao quát tất cả các bình luận đánh giá sau khi
trải nghiệm trên website agoda.

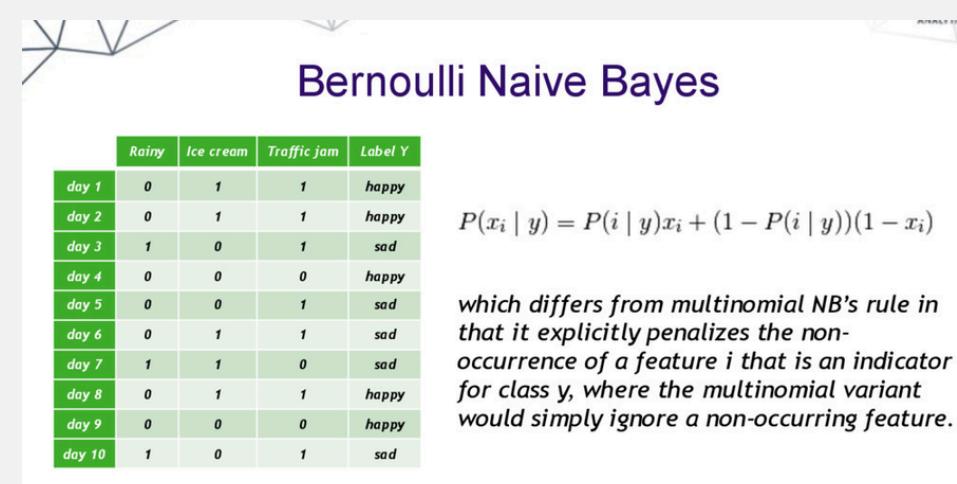




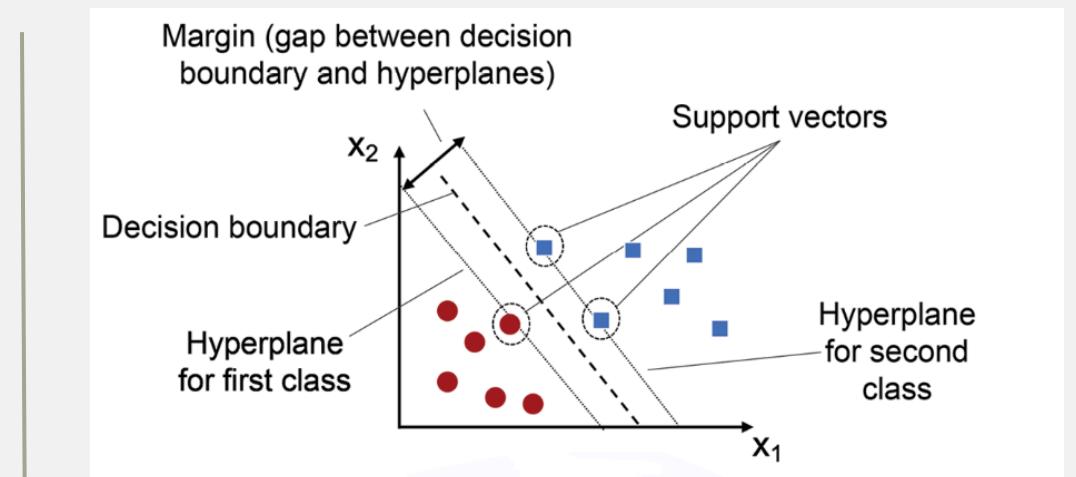
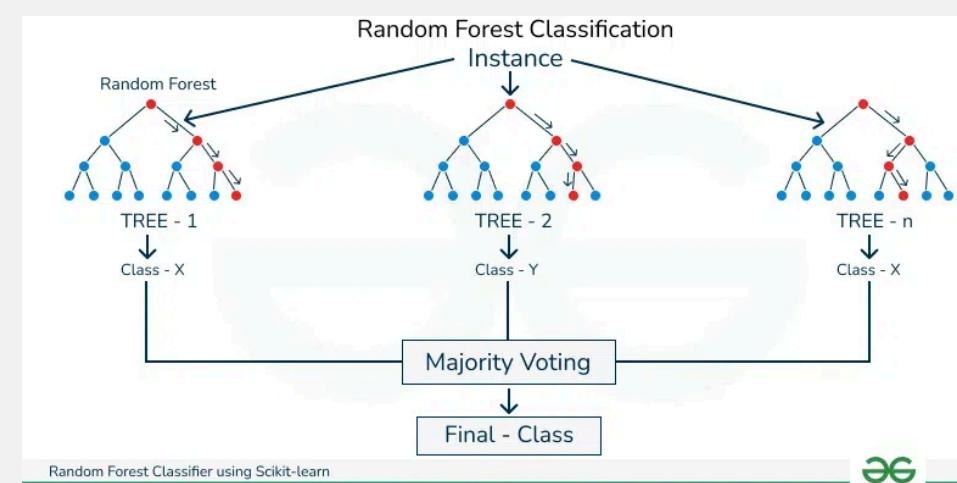
Natural Language Processing



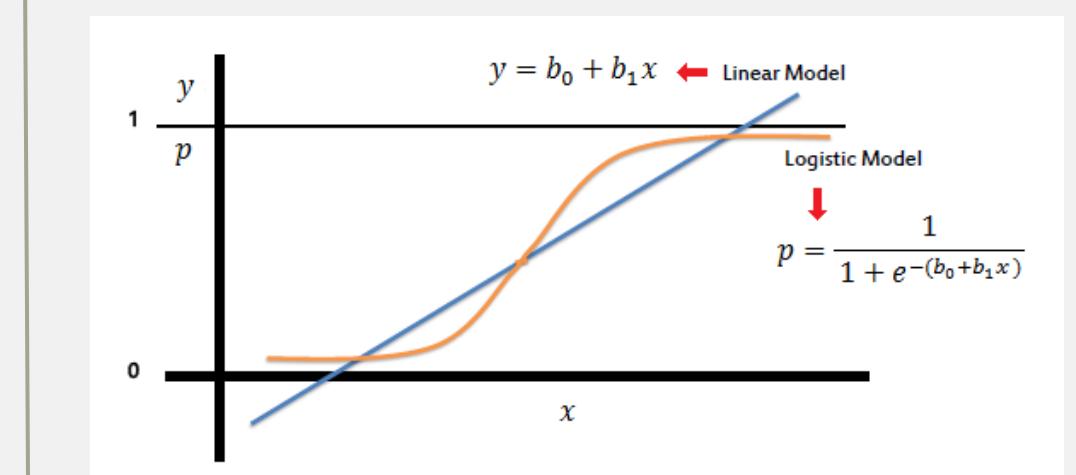
Bernoulli Naive Bayes



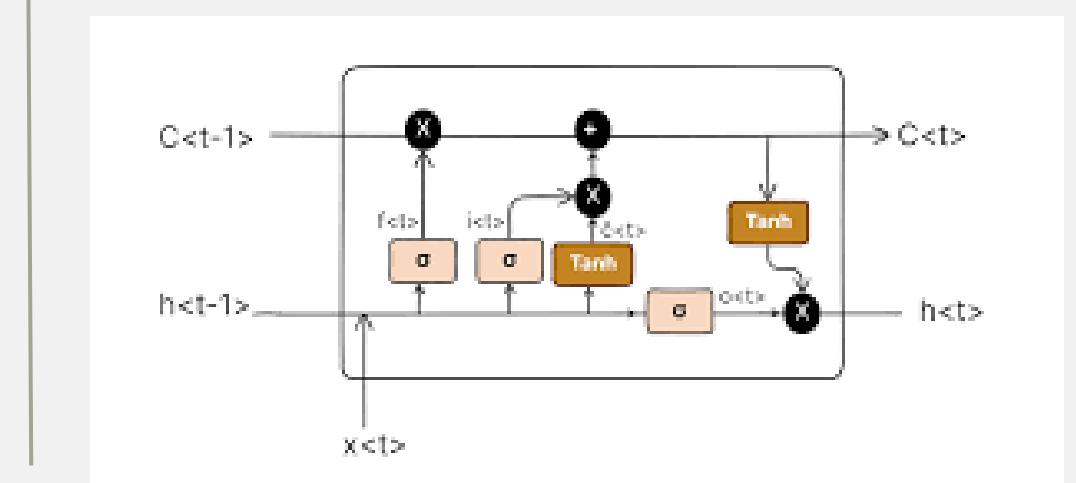
Random Forest



Support Vector Classifier



Logistic Regression



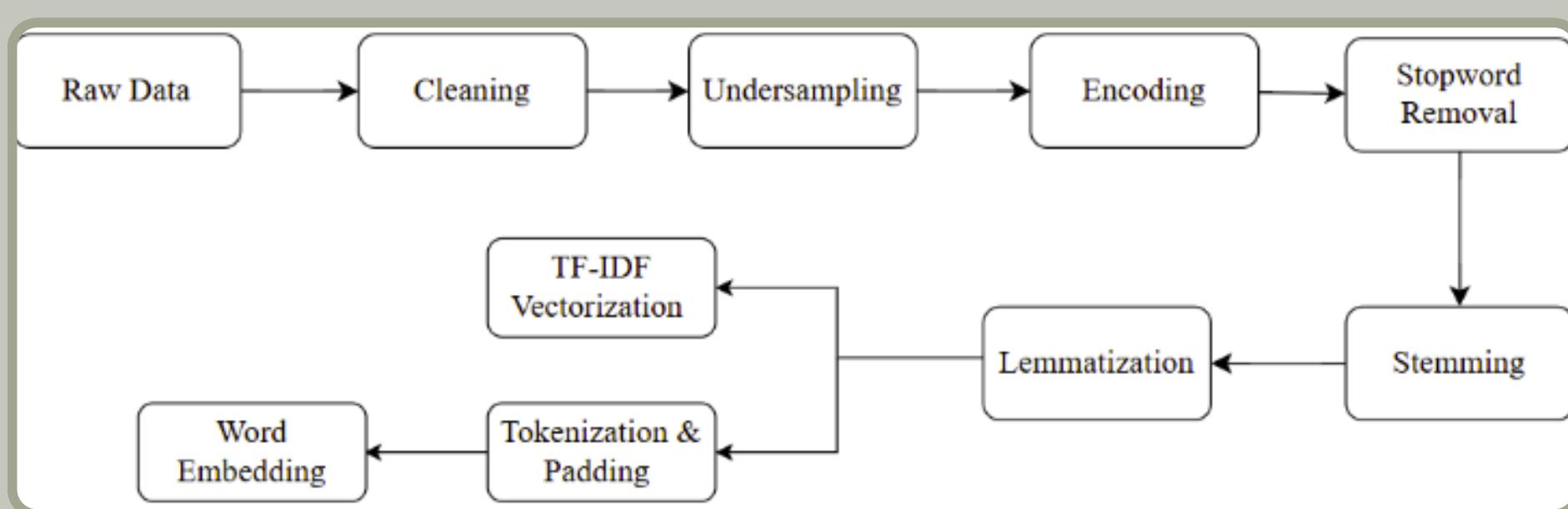
Long Short-Term Memory Networks



PHƯƠNG PHÁP ĐỀ XUẤT



Tiền xử lý dữ liệu



Hình: Quy trình các bước tiền xử lý dữ liệu

01 NLTK and SpaCy

NLTK: Toàn diện, linh hoạt, nhiều công cụ NLP (tokenization, stemming, POS tagging, parsing).

SpaCy: Hiệu suất cao, tối ưu cho dữ liệu lớn và ứng dụng thời gian thực (NER, dependency parsing).

Raw Data

Cleaning

Undersampling

Encoding

Stopword Removal

TF-IDF Vectorization

Lemmatization

Stemming

Word Embedding

Tokenization & Padding

01

NLTK and SpaCy

NLTK: Toàn diện, linh hoạt, nhiều công cụ NLP (tokenization, stemming, POS tagging, parsing).

SpaCy: Hiệu suất cao, tối ưu cho dữ liệu lớn và ứng dụng thời gian thực (NER, dependency parsing).

02

Undersampling



Kỹ thuật cân bằng dữ liệu bằng cách giảm số lượng mẫu ở lớp đa số.



Ứng dụng: Loại bỏ bớt mẫu từ nhóm có Review_Rating 5, 4, 3 để cân bằng tập dữ liệu.

03

Encoding



Label Encoding: Chuyển dữ liệu danh mục thành số nguyên

“Good” -> 1; “Bad” -> 0

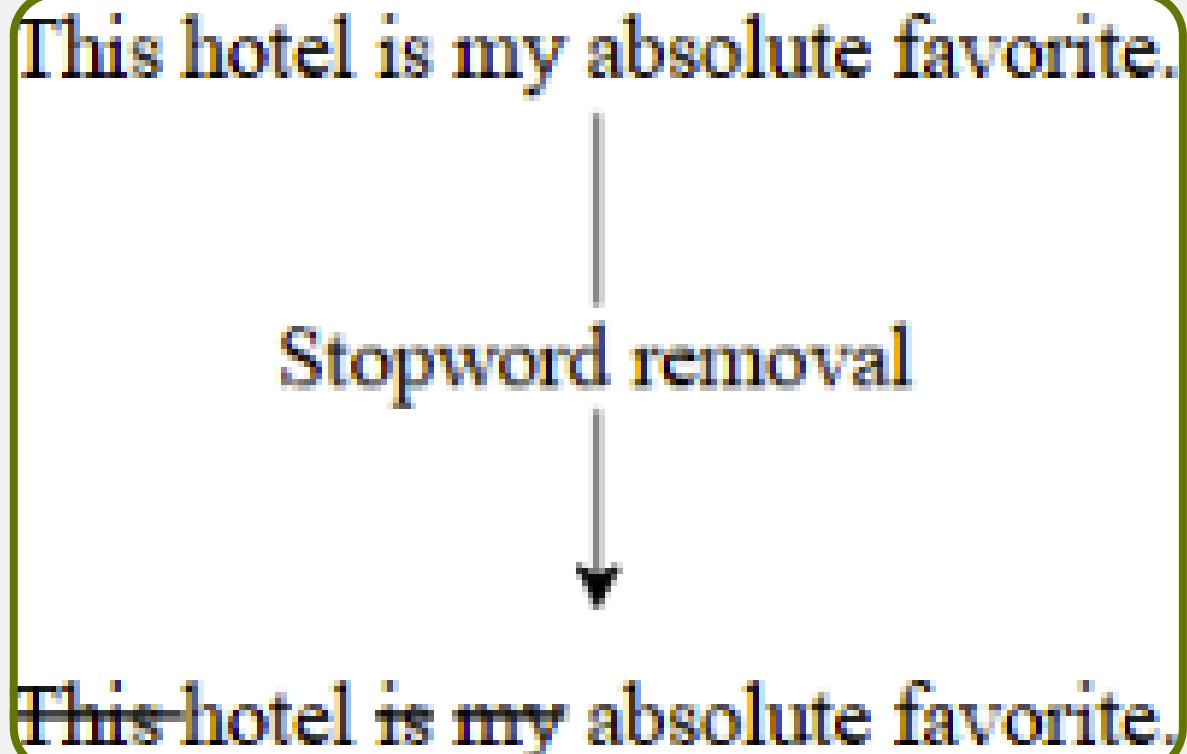
Tiền xử lý dữ liệu

04

Stopword Removal



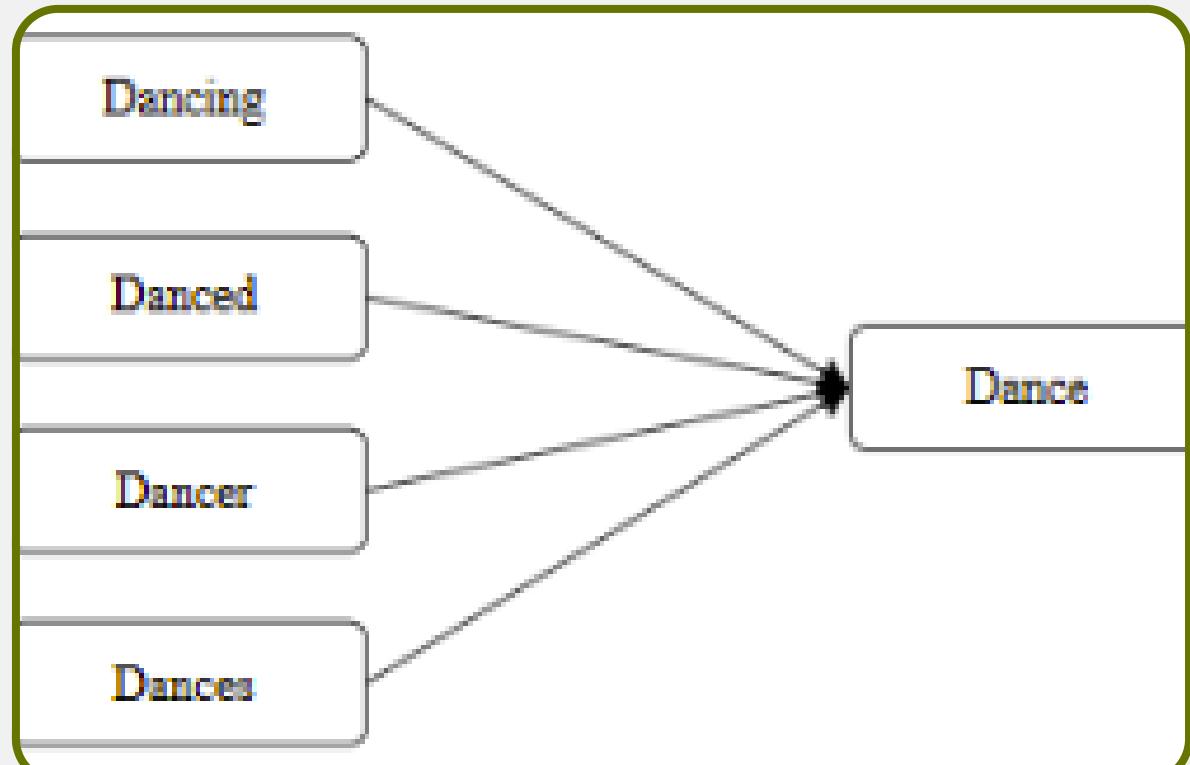
Loại bỏ từ không quan trọng trong NLP (vd: "the", "is", "and").



05

Stemming

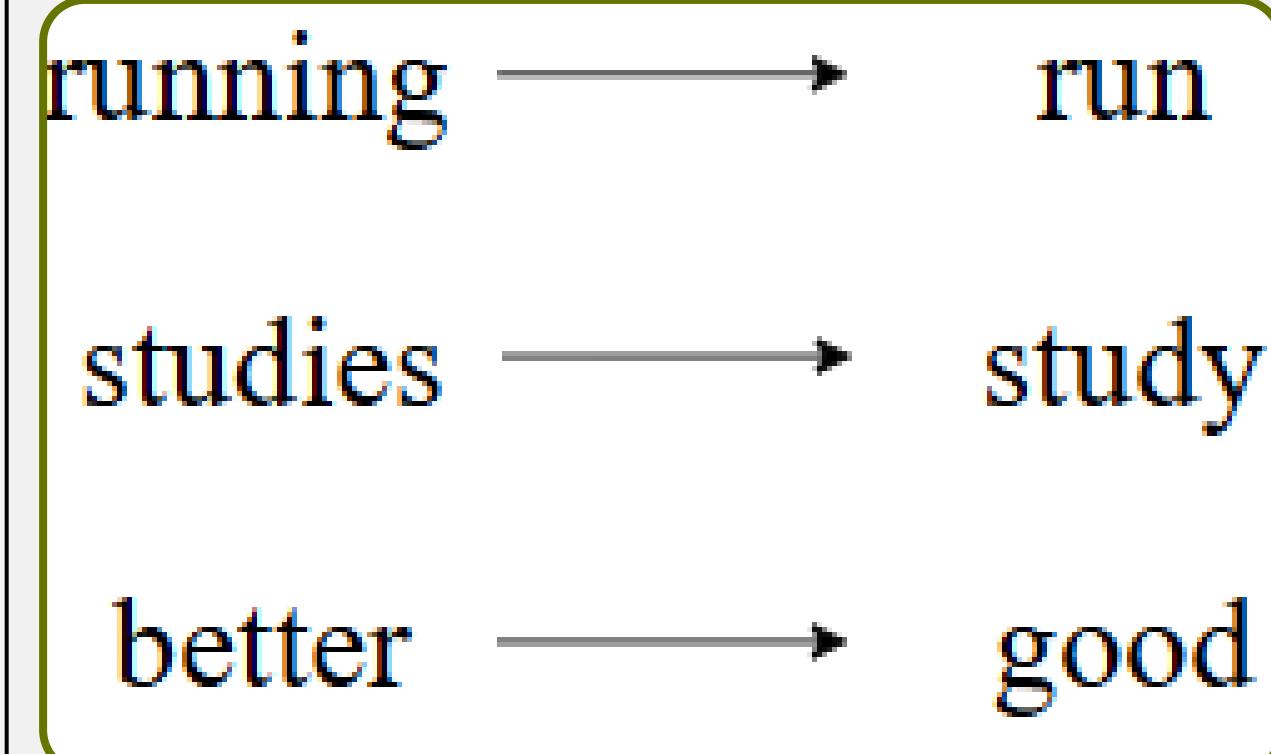
Đưa từ về dạng gốc bằng cách loại bỏ tiền tố/hậu tố.



06

Lemmatization

Đưa từ về dạng gốc (lemma) dựa trên ngữ cảnh & loại từ.



Tiền xử lý dữ liệu

07

TF-IDF Vectorization

Chuyển văn bản thành số để xử lý trong NLP.

TF (Term Frequency): Tần suất từ trong tài liệu; IDF: Mức độ quan trọng của từ trong toàn bộ tập dữ liệu.

Dùng trong mô hình: BernoulliNB, Logistic Regression, Random Forest, SVC.

08

Tokenization & Padding

Tokenization: Chuyển văn bản thành danh sách token (từ hoặc ký tự).

➡ Ví dụ: "Hôm nay trời đẹp quá!" → ["Hôm", "nay", "trời", "đẹp", "quá", "!"]

Padding: Đảm bảo đầu vào có độ dài cố định để mô hình LSTM xử lý dễ dàng, tránh lỗi

09

Word Embedding

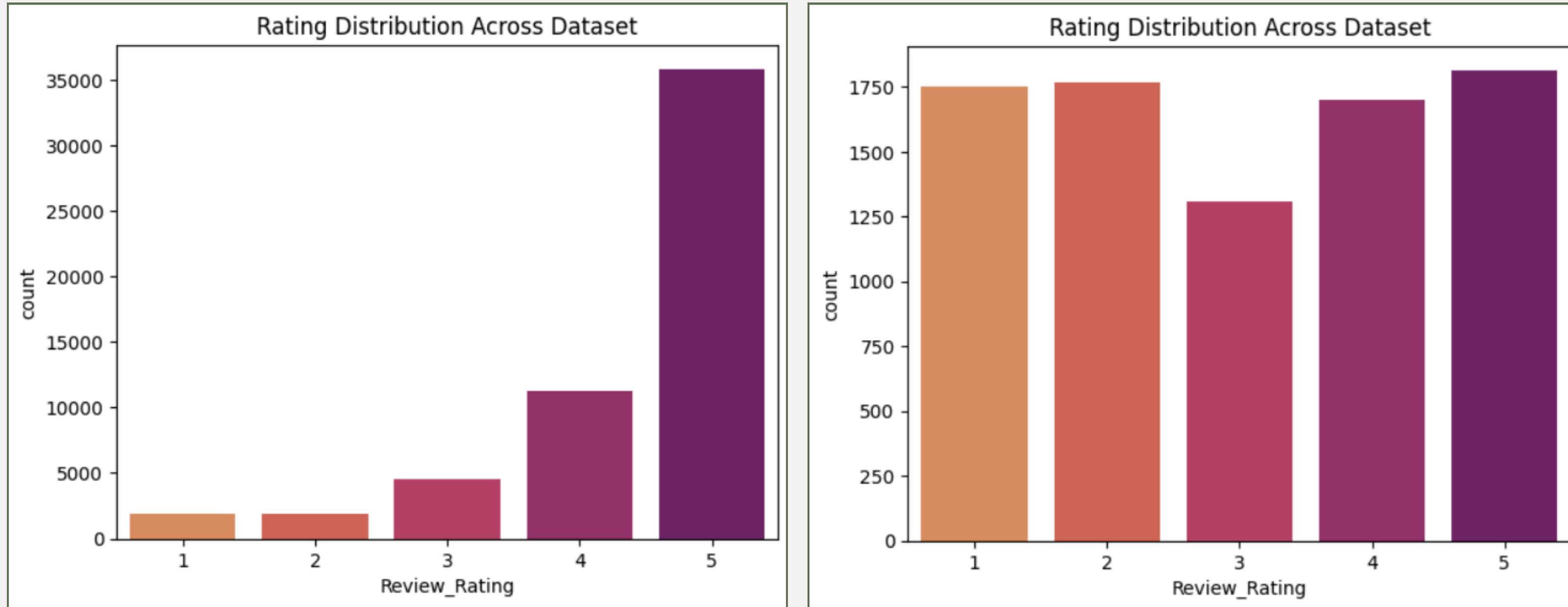
Biểu diễn từ trong không gian vector, thể hiện mối quan hệ ngữ nghĩa.

➡ Ví dụ: "Táo" và "Xoài" có vị trí gần nhau trong câu "Hôm nay ăn táo/xoài"

Chuyển token thành vector số trước khi đưa vào LSTM

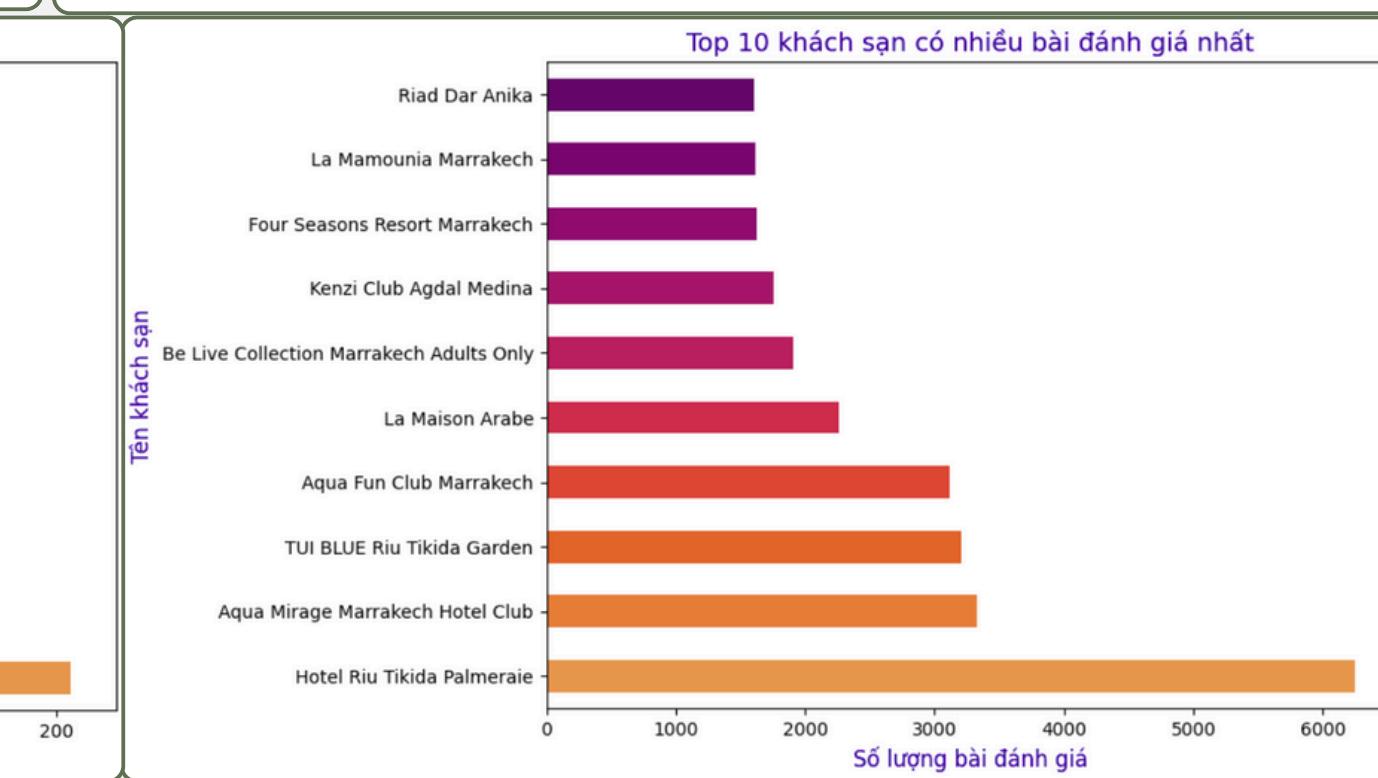
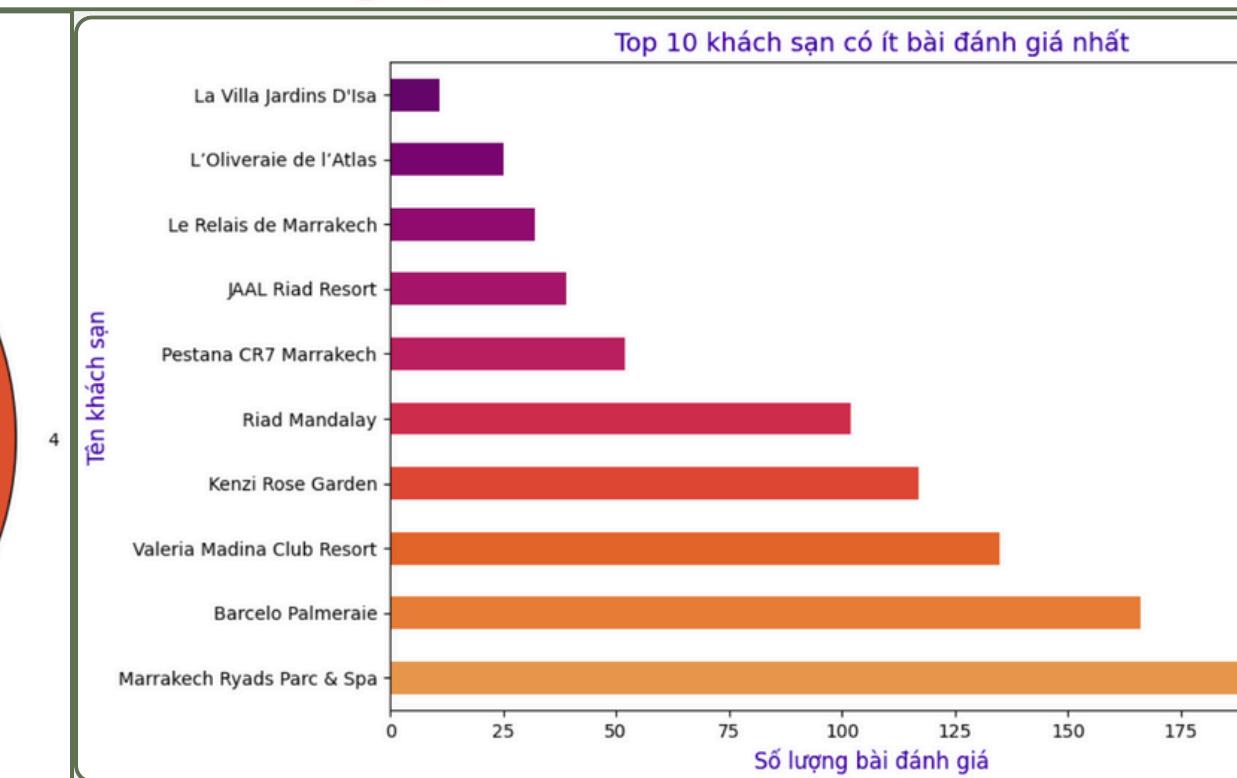
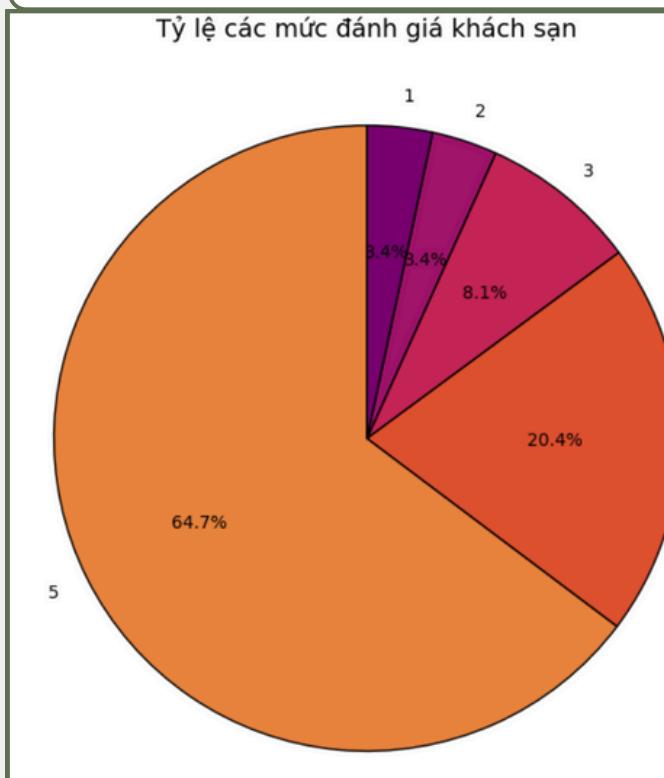
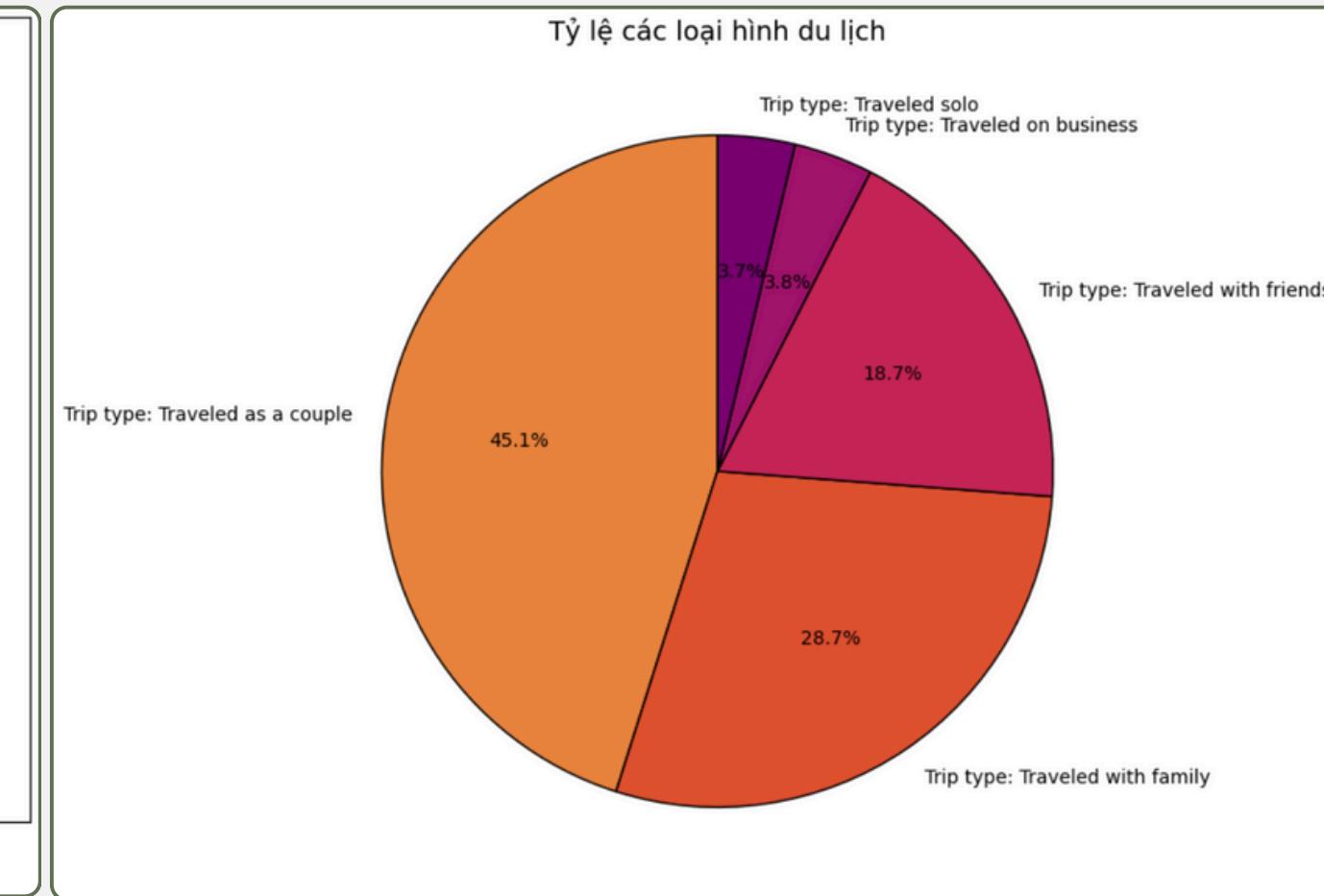
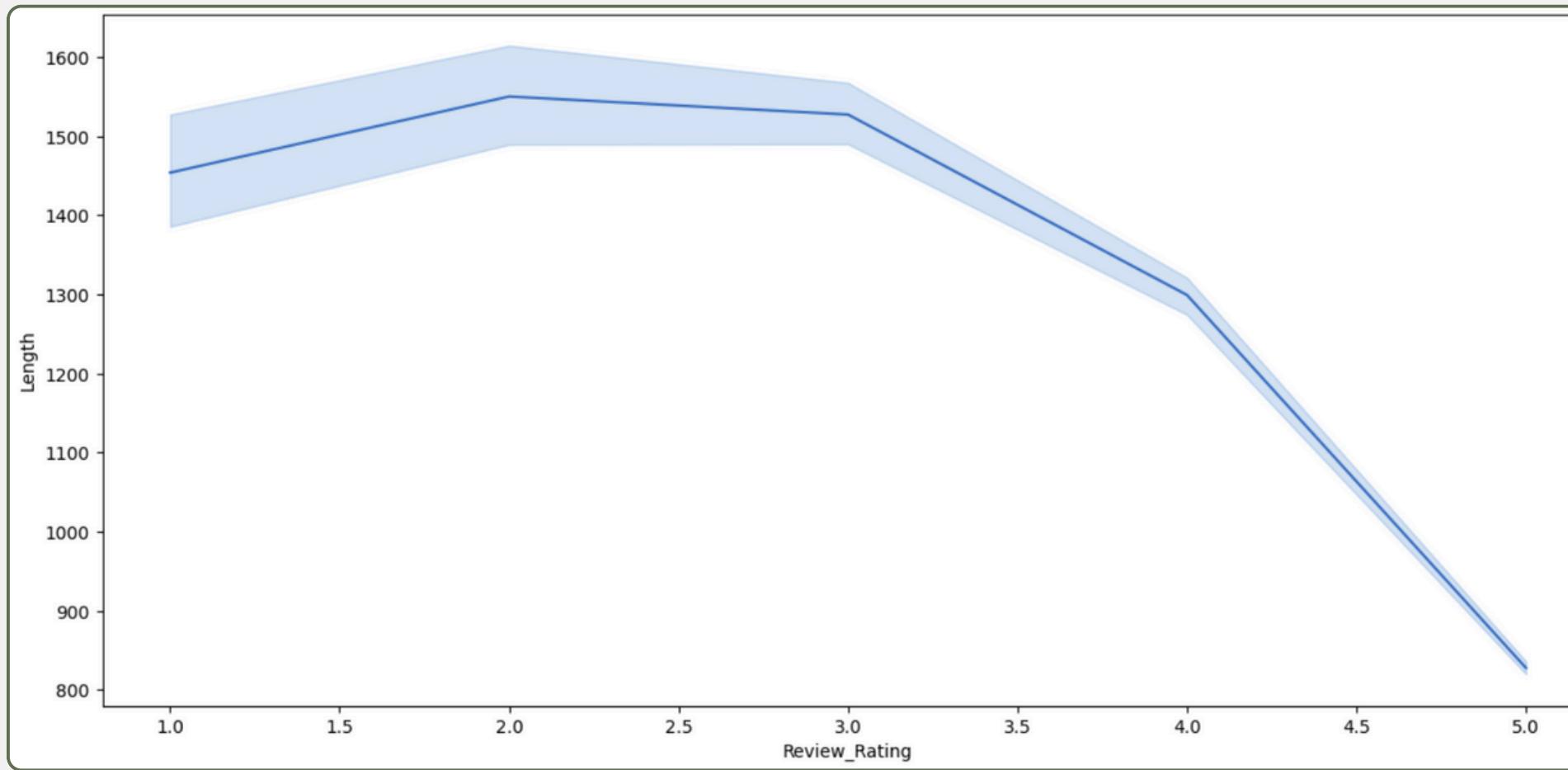


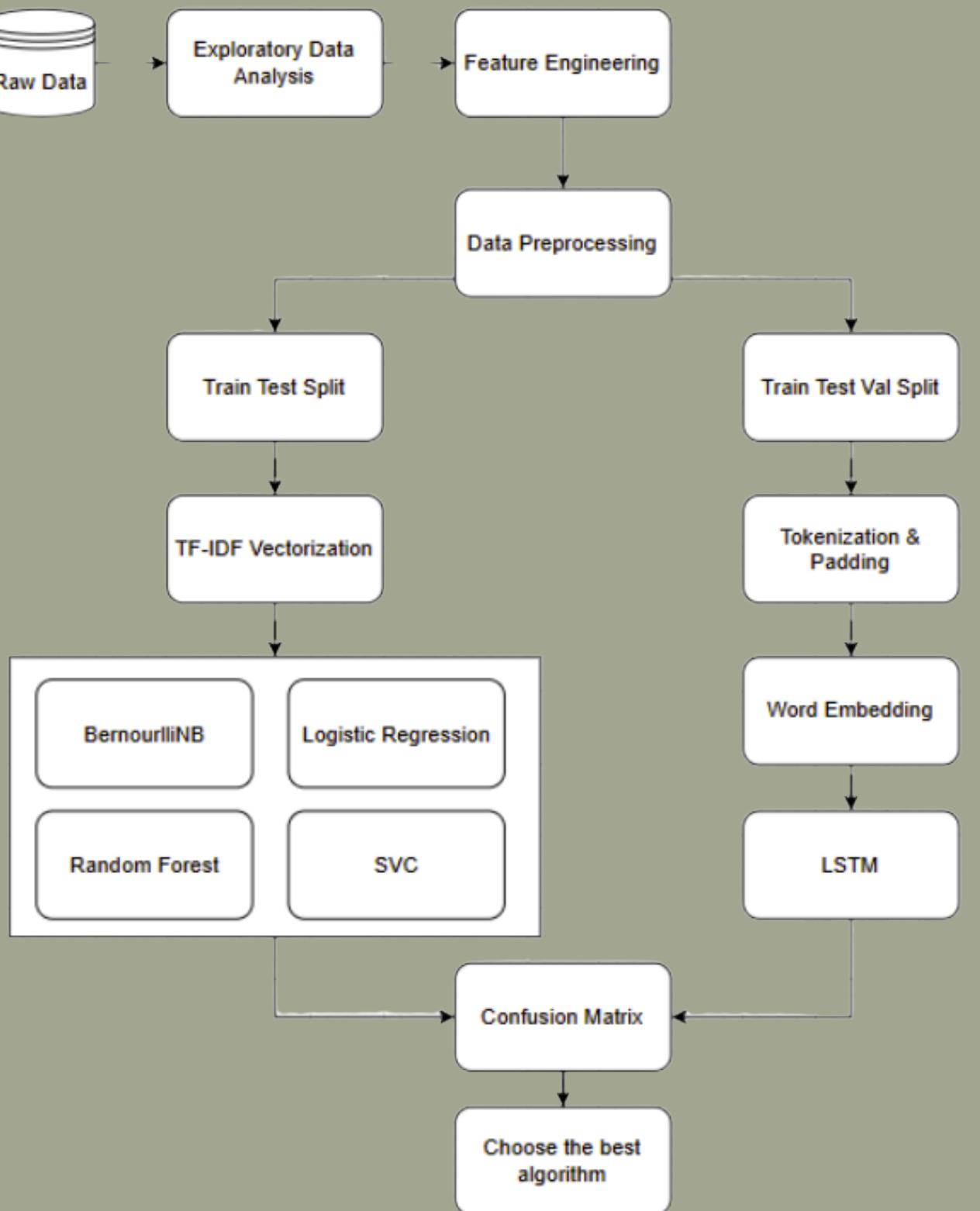
Trực quan hóa dữ liệu



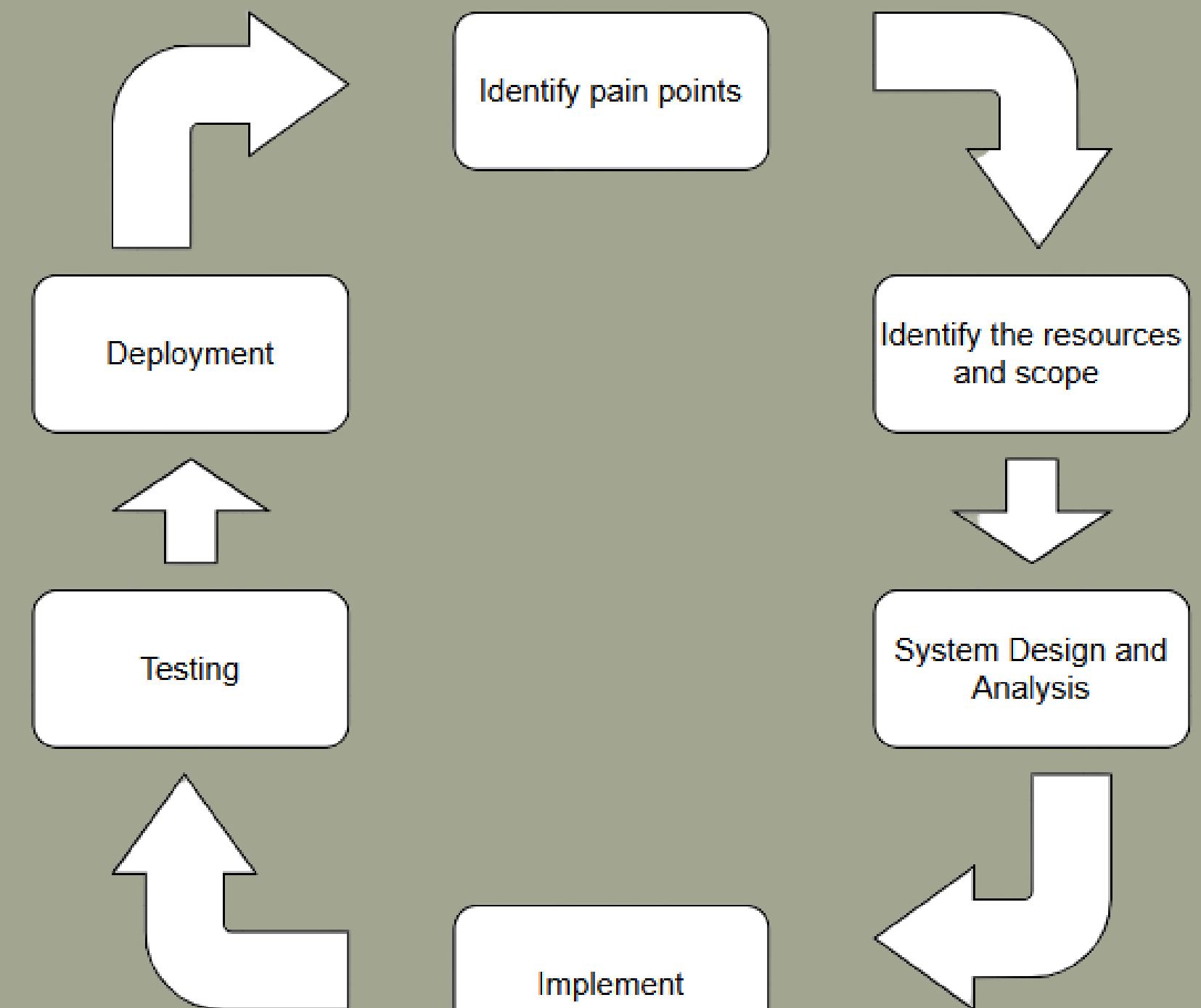
Phân bổ rating trước và sau khi xử lý data imbalanced

Trực quan hóa dữ liệu





Quy trình xây dựng mô hình NLP



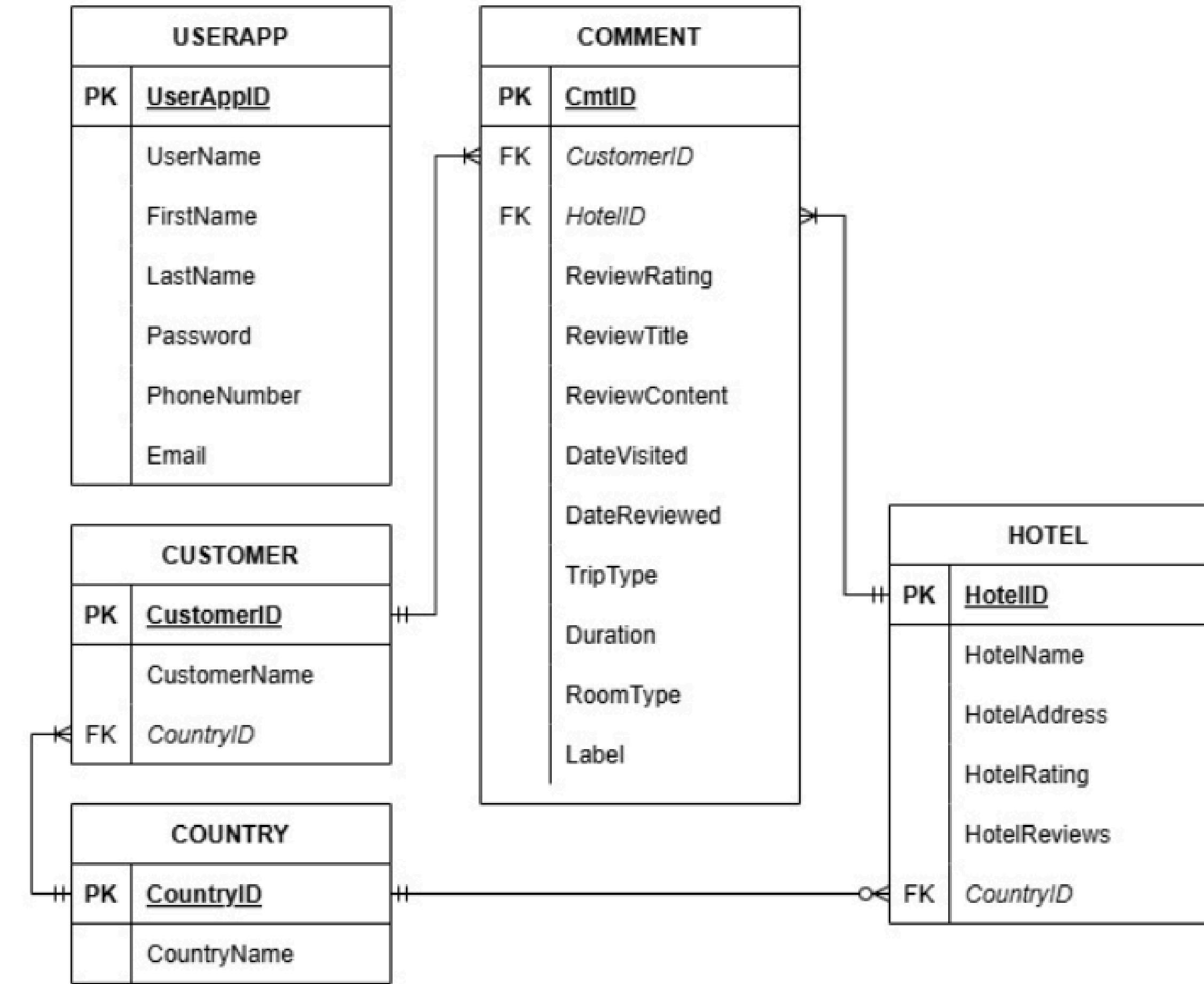
Quy trình phát triển ứng dụng



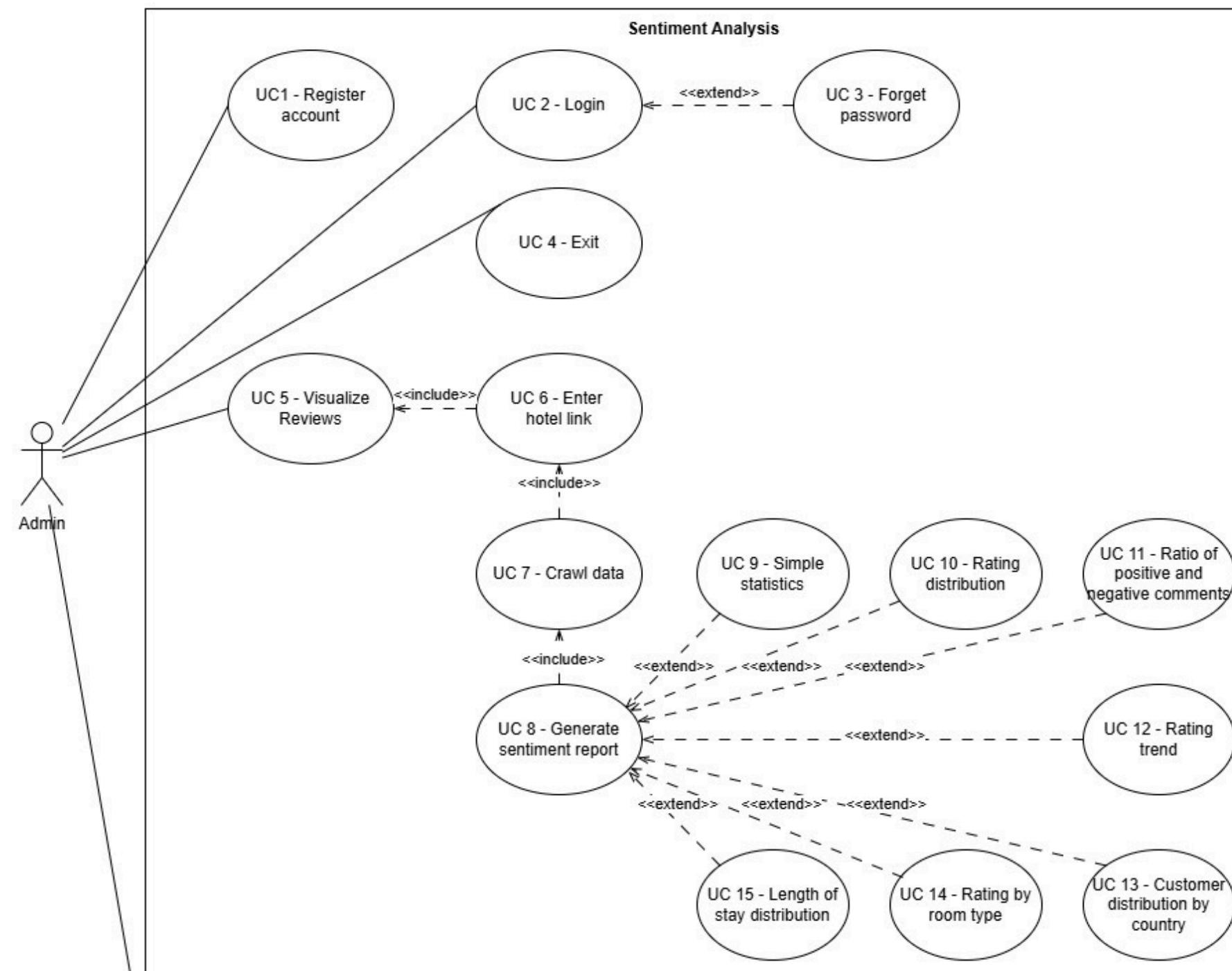
MÔ HÌNH VÀ QUY TRÌNH THỰC HIỆN



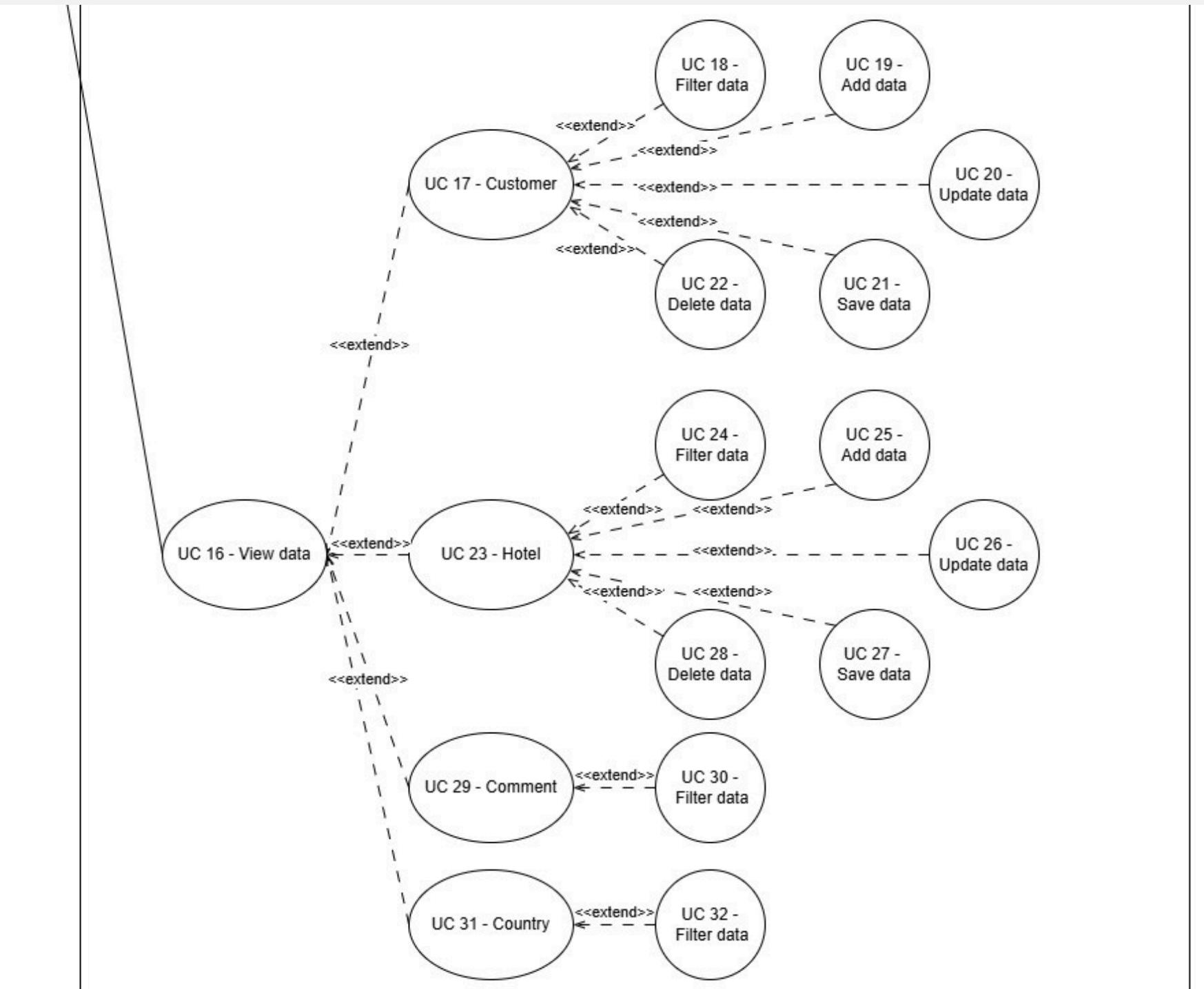
Thiết kế cơ sở dữ liệu



Database ERD



Use case diagram





KẾT QUẢ THỰC NGHIỆM



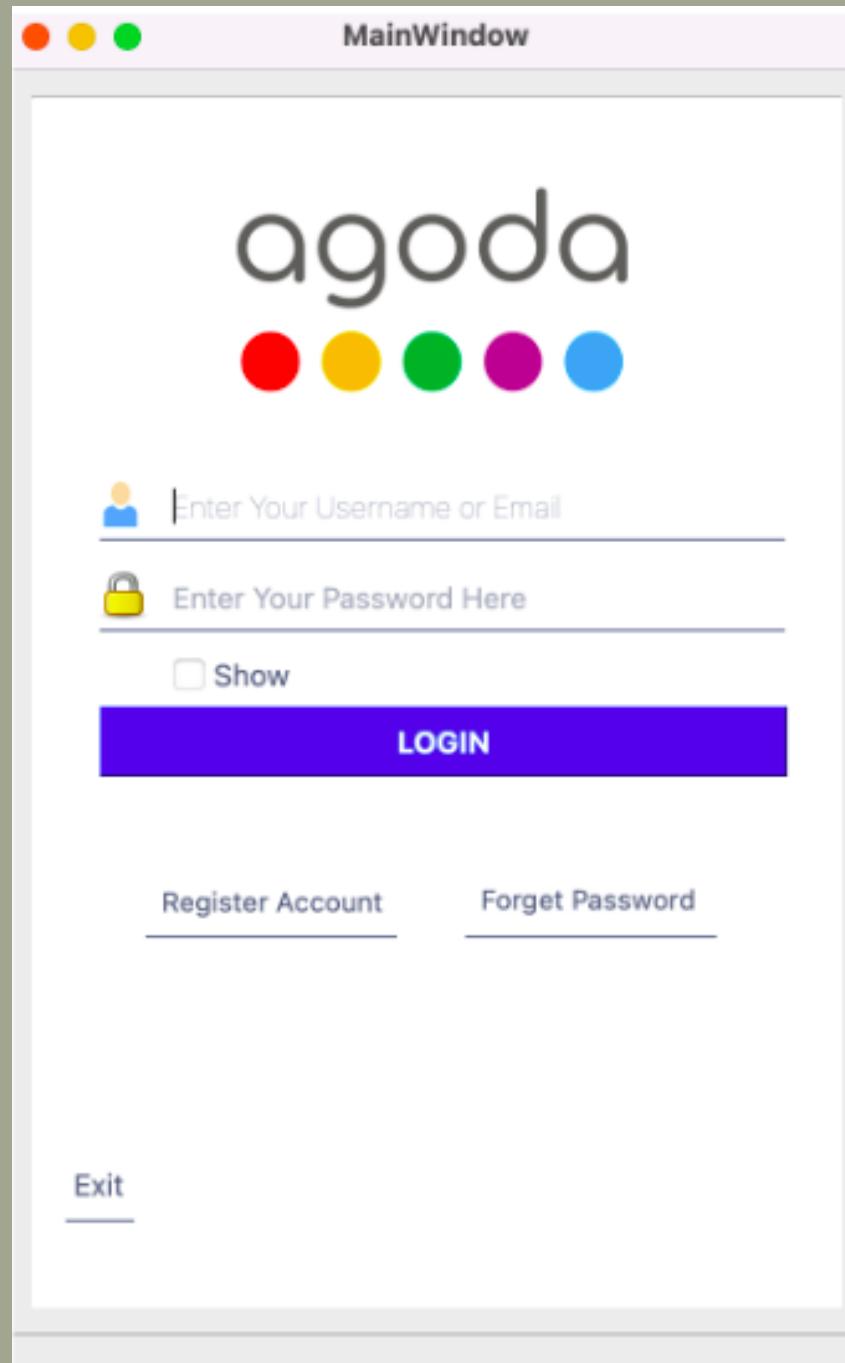
Đánh giá mô hình

	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
4	LSTM	0.900539	0.909785	0.847578	0.877581

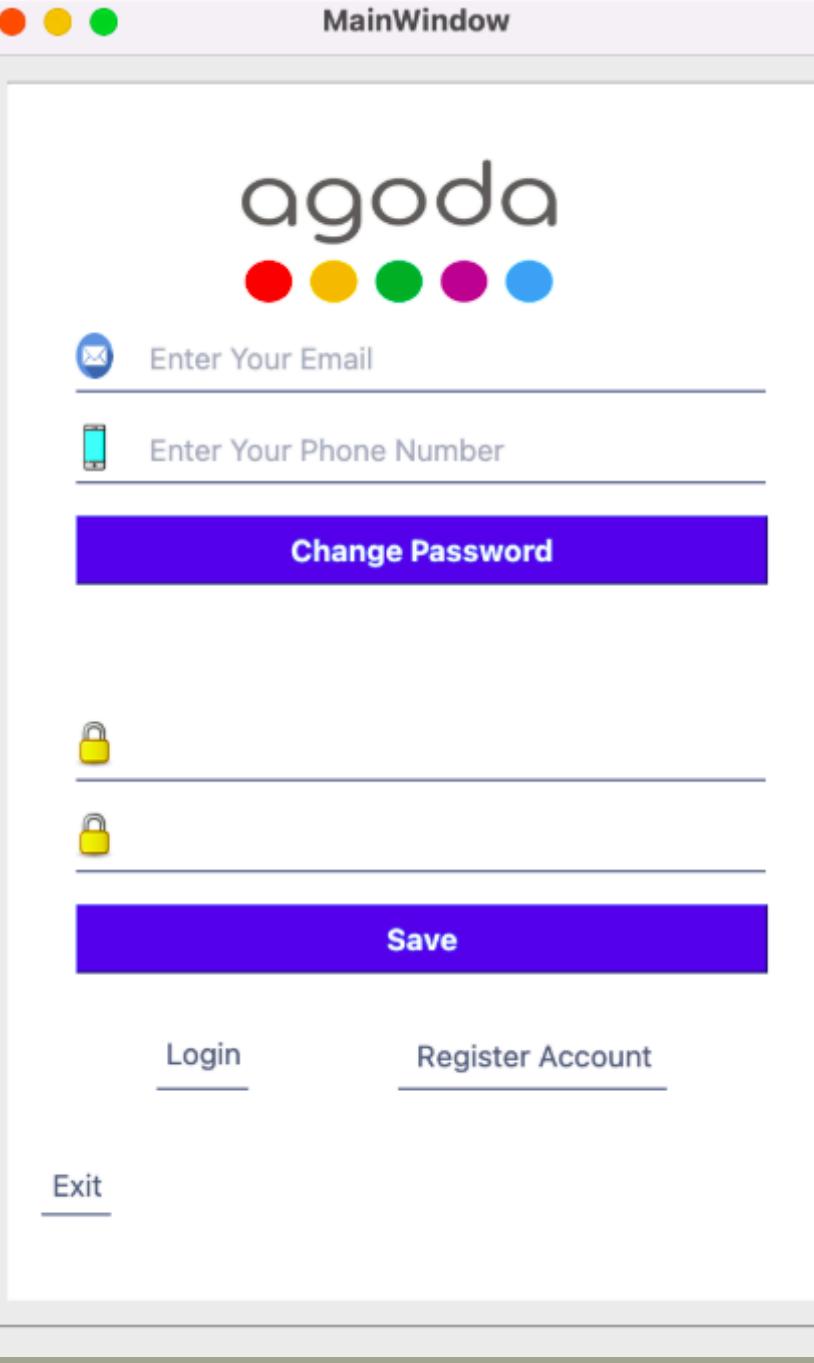


Mô hình LSTM là lựa chọn tối ưu cho bài toán phân tích cảm xúc

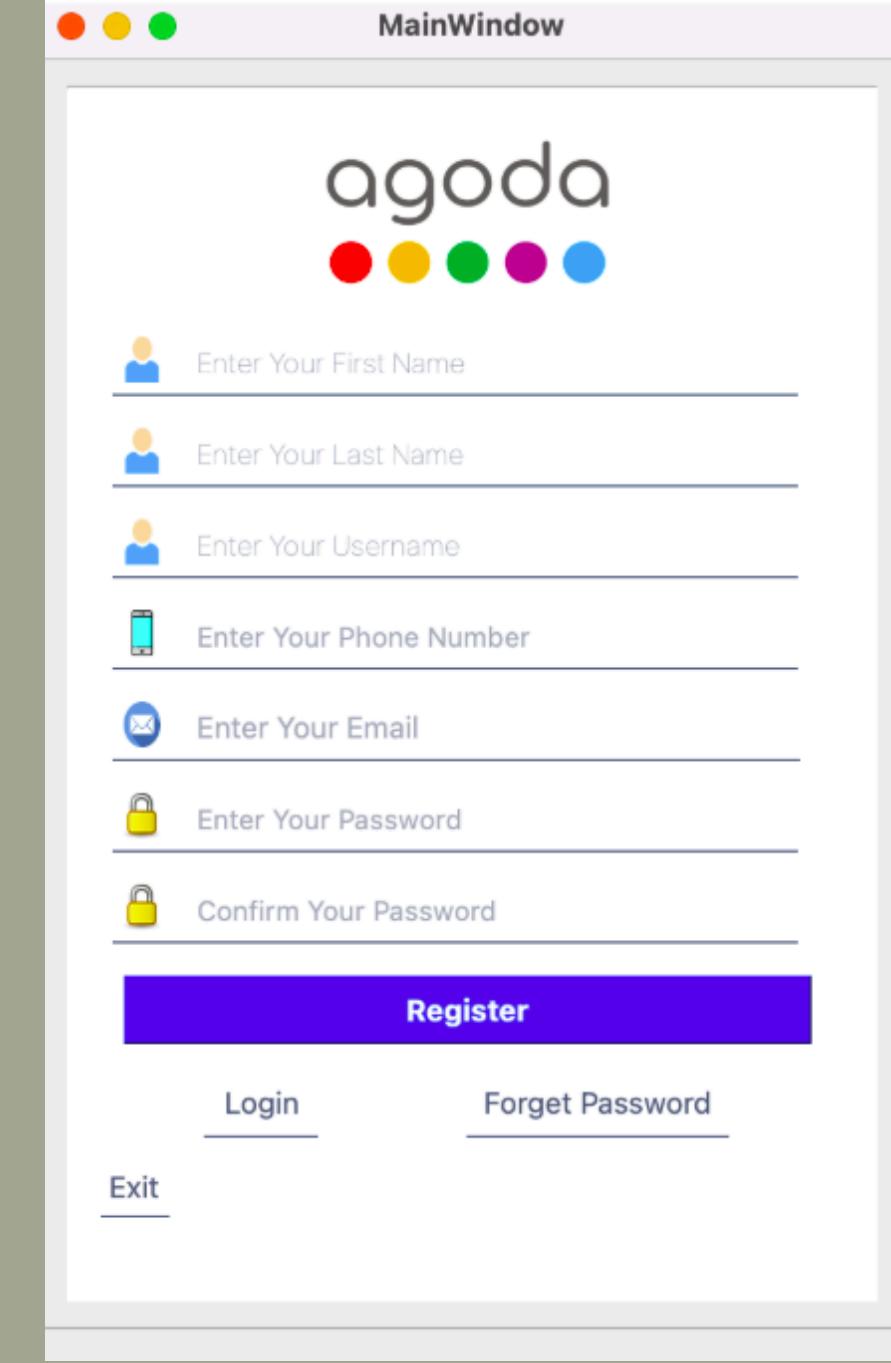
Giao diện người dùng



Màn hình đăng nhập

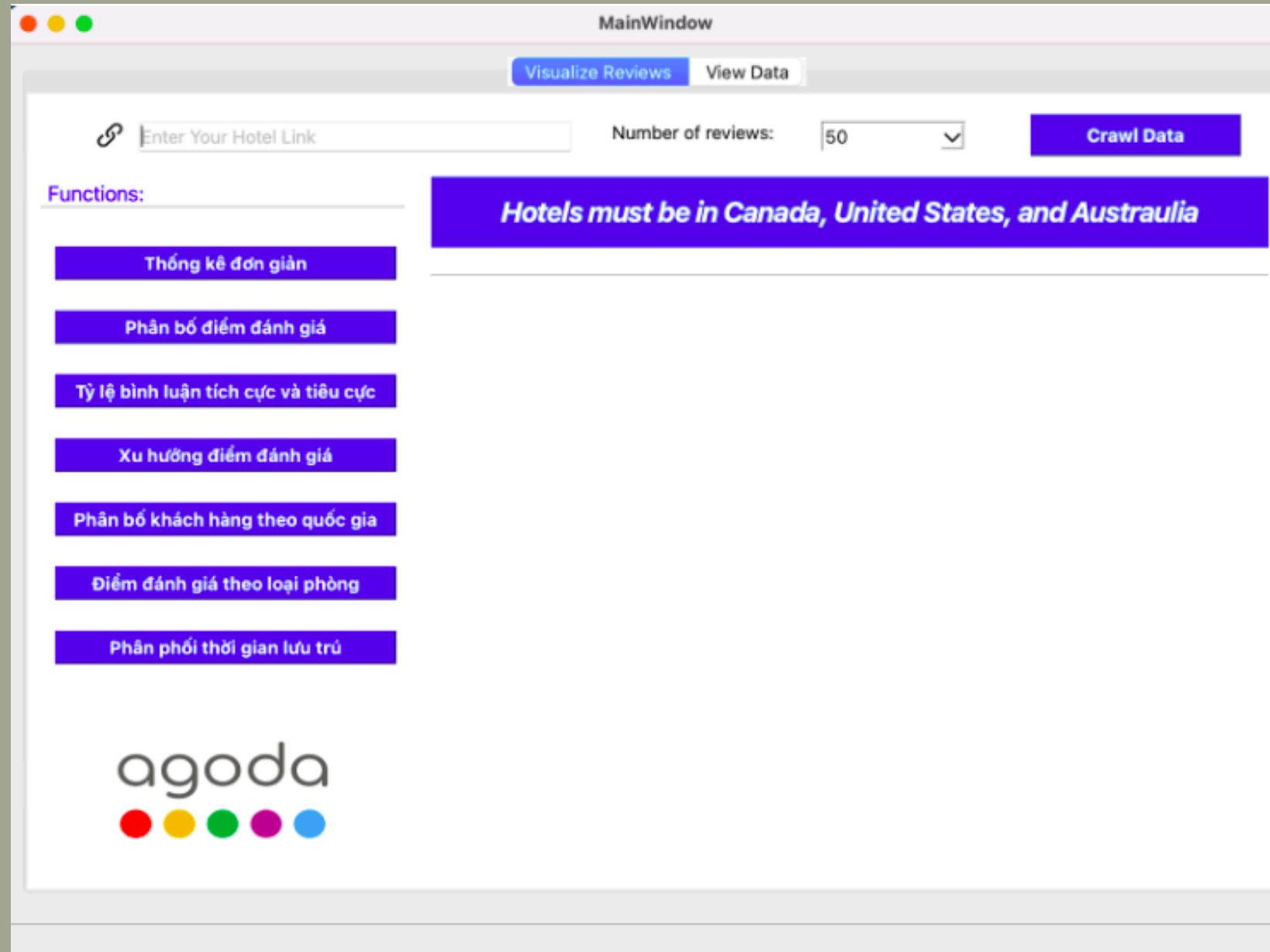


Màn hình quên mật khẩu

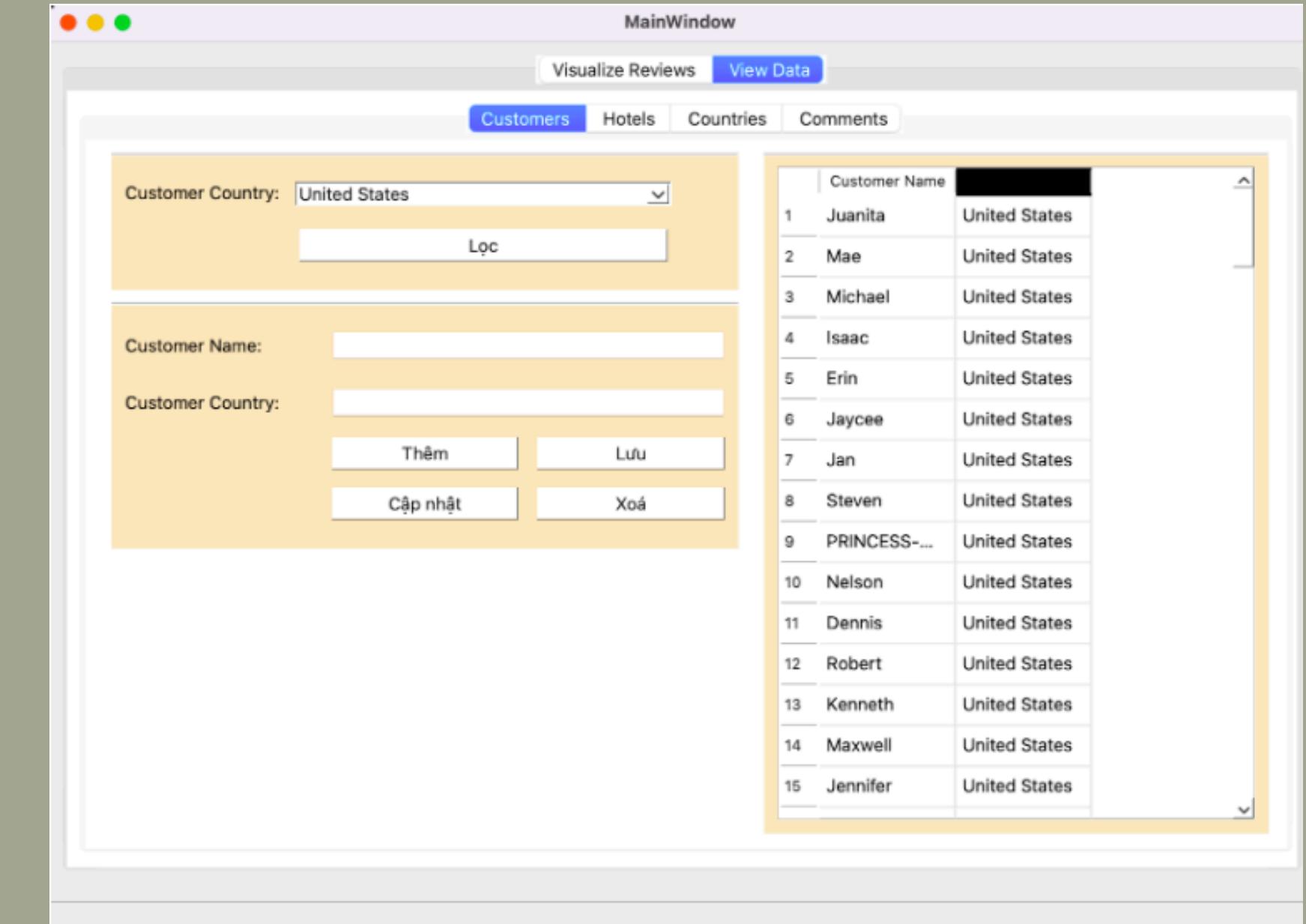


Màn hình đăng ký tài khoản

Giao diện người dùng



Trực quan hóa đánh giá



Hiển thị dữ liệu khách hàng



KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN





Sentiment Analysis App

-  **Mục tiêu:** Hỗ trợ doanh nghiệp khách sạn hiểu phản hồi khách hàng.
-  **Công nghệ:** NLP + ML (Naïve Bayes, Random Forest, Logistic Regression, SVC) & Deep Learning (LSTM).
-  **Tích hợp:** CSDL + giao diện trực quan giúp phân tích & ra quyết định hiệu quả.





Hạn chế và Hướng phát triển



01

Chưa hiểu ngữ
nghĩa phức tạp
(ẩn dụ, mỉa mai)
→ Tích hợp mô
hình BERT/GPT
để xử lý tốt hơn.

02

Chưa hỗ trợ đa
ngôn ngữ
→ Mở rộng để
phân tích cảm
xúc nhiều ngôn
ngữ khác nhau.

03

Tốc độ & tài
nguyên chưa tối
ưu → Cải thiện
bằng giảm kích
thước mô hình,
song song hóa,
dùng GPU.





Thank You