Đề tài: Predicting Customer Emotions From Product Reviews

Thành viên:

- Nguyễn Lư Hồng Phương - 2251010077

- Nguyễn Thị Mai - 2251010062

Mục lục

[I. Giới thiệu đề tài 1](#_Toc16576)

[1. Mục tiêu 1](#_Toc8402)

[2. Tính ứng dụng 1](#_Toc25976)

[I. Phương pháp áp dụng 2](#_Toc2721)

[1. Crawl Data: 2](#_Toc7368)

[- Nguồn: 2](#_Toc4253)

[2. Processing Data: 2](#_Toc26800)

[2.1. Clean data 2](#_Toc28271)

[2.2. Normalize data 2](#_Toc31033)

[2.3. Analyse data 2](#_Toc29435)

[2.4. Balance data. 3](#_Toc11464)

[3. Text Feature: 3](#_Toc27093)

[4. Machine Learning model 3](#_Toc5186)

[4.2. Naive Bayes 3](#_Toc19818)

[4.3. Random Forest 4](#_Toc8469)

[4.4. SVM 4](#_Toc28510)

[4.5. Logistic Regression 4](#_Toc5268)

[II. So sánh và kết luận. 5](#_Toc28912)

[1. Kết luận 5](#_Toc17238)

[2. Hướng phát triển: 5](#_Toc29536)

[III. Ứng dụng 5](#_Toc5243)

# Giới thiệu đề tài

## Mục tiêu

* Mục tiêu: Nhằm phân tích cảm xúc của người dùng thông qua bình luận sản phẩm. Xây dựng mô hình Machine Learning để phân loại cảm xúc (Cực kì hài lòng và không hài lòng) dựa trên dữ liệu được thu thập từ trang thương mại điện tử Tiki.

## Tính ứng dụng

* Tính cấp thiết của đề tài thể hiện qua nhu cầu nâng cao chất lượng, uy tín của doanh nghiệp trong việc: Hiểu rỏ tâm lý khách hàng, từ đó có chiến lược cải tiến sản phẩm và dịch vụ phù hợp. Tự động hóa quá trình đọc và giải quyết ý kiến khách hàng giúp tiết kiệm thời gian và nhân lực. Góp phân nâng cao trải nghiệm của người dùng trên nền tảng thương mại điển tự với nhiều mặt hàng sản phẩm khác nhau.

# Phương pháp áp dụng

## Crawl Data:

- Nguồn: Dữ liệu lấy từ trang thương mại điện tử Tiki

- Phương pháp: Sử dụng kỹ thuật **web crawling thông qua API**. Cụ thể, dùng thư viện requests trong Python để gửi các truy vấn đến API của Tiki. Với tổng 26 danh mục và tất cả sản phẩm thuộc danh mục đó, nên tập dữ liệu ban đầu được thu thập là 69864 dữ liệu.

#### Bảng Tiki\_Comments.csv

- Gồm 6 thuộc tính:

+ title: Mức độ đánh giá người dùng.

+ content: Nội dung đánh giá

+ thank\_count: Lượt thích bình luận

+ customer\_id: Mã khách hàng bình luận

+ rating: Sao đánh giá sản phẩm

+ customer\_name: Tên khách hàng bình luận

## Processing Data:

### Clean data

Quy trình làm sạch dữ liệu gồm các bước chính:

* **Loại bỏ dữ liệu trùng lặp (duplicates) và loại bỏ cột không cần thiết** [‘id’, ‘thank\_count’, ‘customer\_name’, ‘customer\_id’].
* **Xử lý giá trị thiếu** (missing value): Loại bỏ những dòng bị thiếu ở cột [‘content’].
* **Không loại bỏ stop words:** Trong bài toán phân tích cảm xúc, stop words (những từ phổ biến như "là", "rất", "không", "có",...) có thể đóng vai trò quan trọng trong việc diễn đạt sắc thái và thái độ cảm xúc của người viết.

+ Ví dụ: Câu "Tôi không hài lòng với dịch vụ." mang cảm xúc tiêu cực, và nếu loại bỏ từ "không", mô hình có thể hiểu sai ý nghĩa. Hoặc trong câu "Sản phẩm rất tốt và tôi rất hài lòng.", từ "rất" giúp nhấn mạnh mức độ cảm xúc tích cực.

=> Vì lý do đó, nhóm quyết định giữ lại stop words để đảm bảo mô hình học được ngữ cảnh và sắc thái đầy đủ của văn bản, từ đó cải thiện độ chính xác trong việc nhận diện cảm xúc.

### Normalize data

* Loại bỏ biểu tượng cảm xúc (emoji)
* thay thế từ viết tắt bằng dạng đầy đủ
* chuyển toàn bộ văn bản về chữ thường
* rút gọn các ký tự lặp lại từ ba lần trở lên
* xóa ký tự đặc biệt và tách từ (word tokenization)
* Thay thế ký tự ‘\_’ sang ‘ ’ đối với cột [‘title’].

### Analyse data

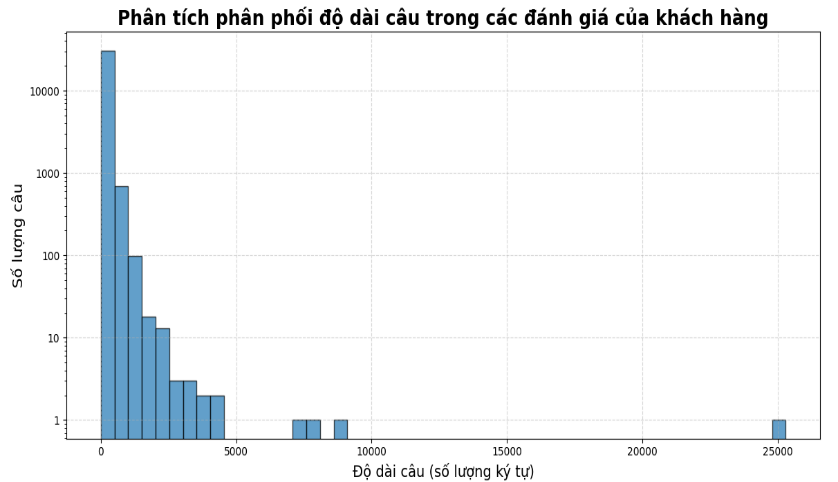
Sau khi thu thập và làm sạch dữ liệu. Nhóm tiến hành phân tích sơ bộ để hiểu rõ hơn về đặc điểm nội dung các bình luận của người dùng.

* + 1. **Mức độ tập trung của các từ.**

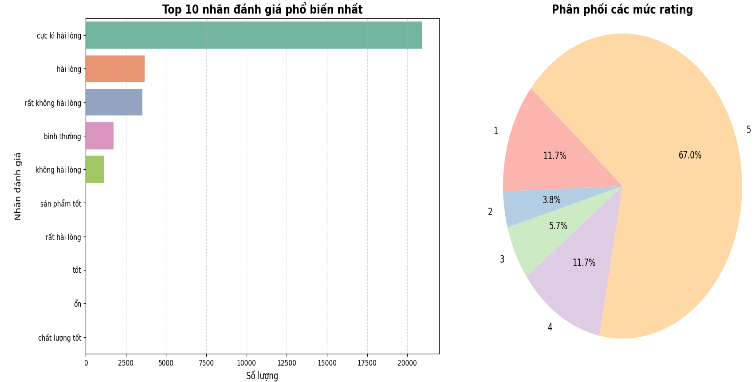
Tổng số từ (tokens) sau khi xử lý là 29 687 từ. Các từ phổ biến trong tập dữ liệu đã có kết quả như sau:

*Most common words: [('ổn', 460), ('tốt', 251), ('đẹp', 76), ('sản\_phẩm tốt', 75), ('good', 73), ('hài\_lòng', 52), ('sản\_phẩm ổn', 44), ('rất tốt', 40), ('rất hài\_lòng', 33), ('giao hàng nhanh', 30), ('tuyệt\_vời', 28), ('chất\_lượng tốt', 27), ('tot', 23), ('ngon', 21), ('hàng ổn', 20)].*

* + 1. **Phân phối độ dài câu trong các đánh giá của khách hàng.**



Từ biểu đồ cho thấy:

* **Phân phối lệch trái**. Số lượng câu dài giảm dần nhanh chóng.
* Cột đầu tiên (0–500 ký tự) có số lượng áp đảo với hơn 10.000 câu.
* Các bình luận của khách hàng **tập trung cao ở những nhận xét ngắn gọn**
  + 1. **Sự phân bố giữa label và rating.**

Nhãn đánh giá sẻ phụ thuộc vào số lượng rating.

Mức rating 5 sao chiếm áp đảo với 67%

* Từ tất cả các phân tích với bộ dữ liệu, nhóm có nhận định với **31678** sau khi làm sạch dữ liệu. Trung bình mỗi bình luận dài khoản 0-500 ký tự với nhiều từ khóa phổ biến như “ổn”, “tốt”, “đẹp”,”sản phẩm tôt”,…

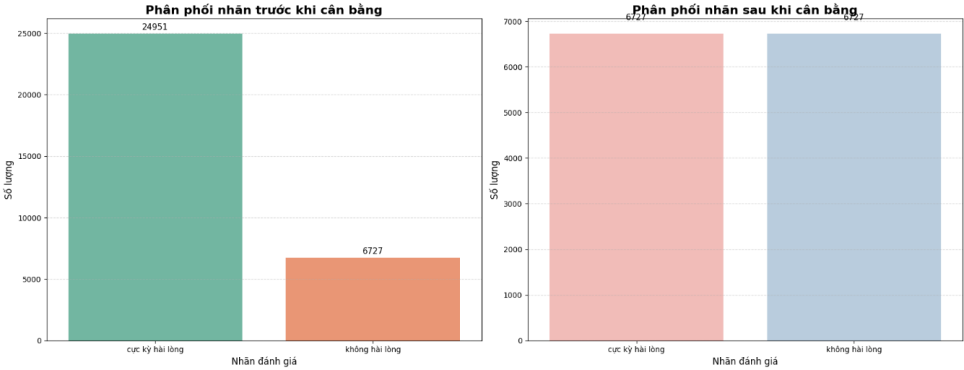
Dữ liệu được gán nhãn cảm xúc theo **5 mức độ rating**:cực kì hài lòng, hài lòng, bình thường, không hài lòng và rất không hài lòng. Kết quả phân bố nhãn đánh giá cho thấy sự mất cân bằng tương đối rõ rệt: 67% rất hài lòng. Đây là yếu tố cần được cân nhắc khi xử lý mất cân bằng dữ liệu.

### Balance data.

- Nhóm tiến hành gôm nhóm từ 5 nhãn tương ứng với với rating thành 2 nhãn chính là nhãn “cực kỳ hài lòng” và “không hài lòng”.

- Vì số lượng tập dữ liệu có nhãn “cực kỳ hài lòng” chiếm số lượng lớn hơn nhãn “Không hài lòng”, dẫn đến hiện tượng lệch mẫu ảnh hưởng tới kết quả mô hình. Nhóm thực hiện cân bằng mẫu bằng việc chọn ngẫu nhiên 6727 mẫu bằng với số lượng mẫu “Không hài lòng”.

- Kết quả thực hiện:



## Text Feature:

- Nhóm sử dụng phương pháp CountVectorize để trích xuất đặc trưng dữ liệu text. Mục tiêu của phương pháp này:

+ Bag-of-Words (BoW): Đếm số lượng mỗi lần từ xuất hiện trong câu.

+ Không quan tâm đến thứ tự từ trong câu và phù hợp với mô hình Machine Learning.

=> Đầu ra: Ma trận đếm từ - Mỗi hàng tương ứng với một câu. Mỗi cột là một đặc trưng của từ đó.

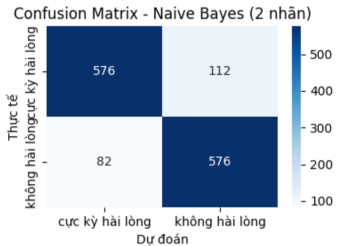
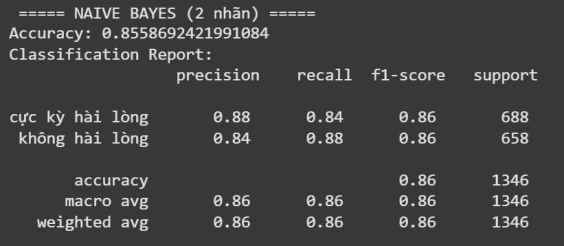
## Machine Learning model

Nhóm tiến hành huấn luyện và đánh giá 4 mô hình học máy phổ biến: Naive Bayes, Random Forest, Logistic Regression và SVM. Nhằm phân loại cảm xúc khách hàng dựa trên bình luận đánh giá sản phẩm.

### Naive Bayes

**- Ưu điểm:** Độ chính xác tốt (**85.6%**) dù mô hình rất đơn giản, dễ cài đặt, huấn luyện nhanh. Precision cao cho nhãn "cực kỳ hài lòng" 88%.

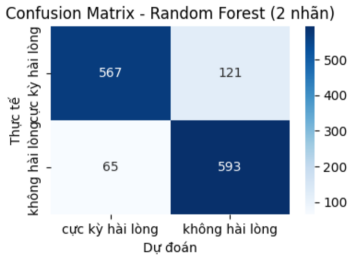
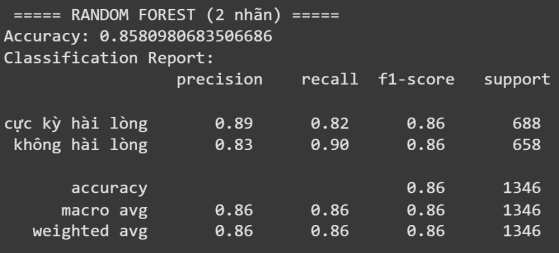
**- Nhược điểm:** Giả định độc lập giữa các từ nên đôi khi không phản ánh đúng ngữ cảnh thực. Số lượng dự đoán sai nhãn “không hài lòng” là **112**, cao hơn so với Logistic.



### Random Forest

**- Ưu điểm:** Có độ chính xác ổn định (**85.8%**), cao hơn Naive Bayes. Recall cao nhất cho nhãn "không hài lòng" là 90%, phù hợp khi cần nhận diện tiêu cực.

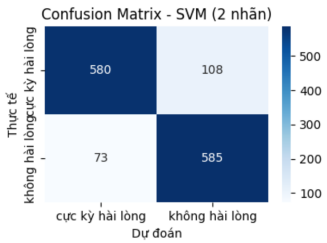
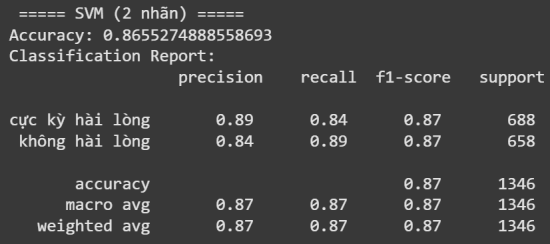
**- Nhược điểm:** Số lượng nhầm lẫn (123 sai nhãn "cực kỳ hài lòng") khá lớn → dễ nhầm cảm xúc tiêu cực so với thực tế.



### SVM

**- Ưu điểm:** Độ chính xác cao (**86.5%**), gần với Logistic. Cân bằng giữa precision và recall → f1-score ở cả hai nhãn đều **0.87**.

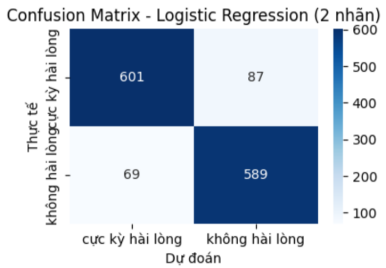
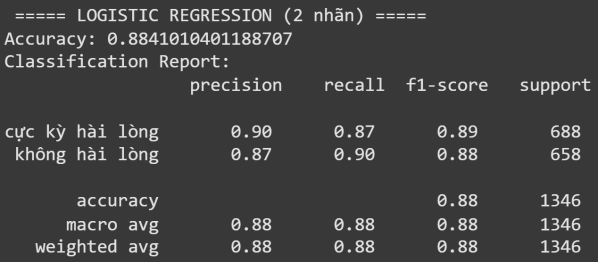
**- Nhược điểm:** Dự đoán sai 108 nhãn “cực kỳ hài lòng” và 73 “không hài lòng” → không vượt trội về độ chính xác. Huấn luyện có thể chậm hơn trên tập dữ liệu lớn, cần điều chỉnh tham số cẩn thận.



### Logistic Regression

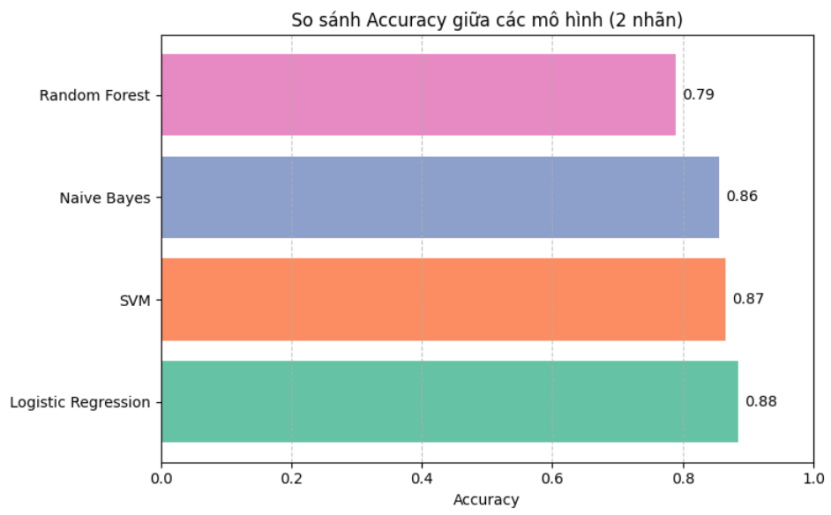
**- Ưu điểm: Độ chính xác cao nhất** (**88.4%**). Precision và recall đều cao, **f1-score** tốt nhất ở cả hai nhãn. Ma trận nhầm lẫn cho thấy ít sai sót hơn: chỉ **87 sai nhãn “cực kỳ hài lòng”** và **69 sai nhãn “không hài lòng”**.

**- Nhược điểm:** Là mô hình tuyến tính → có thể gặp khó khăn với các mối quan hệ phi tuyến tính phức tạp. Phụ thuộc vào việc chuẩn hóa và vector hóa dữ liệu đầu vào.



# So sánh và kết luận.

- Bảng so sánh Accuracy giữa các mô hình với phương pháp trích xuất đặc trưng CountVectorize.



- Bảng phân tích so sánh các mô hình:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mô hình | Đặc điểm | Accuracy | Ưu điểm | Nhược điểm |
| Logistic Regression | Mô hình tuyến tính, hiệu quả cao trên dữ liệu đã vector hóa. | 88.4% | - Accuracy cao nhất. - Cân bằng tốt giữa precision & recall. - Dễ ứng dụng. | - Hạn chế nếu dữ liệu có quan hệ phi tuyến. - Phụ thuộc vào vector hóa chuẩn. |
| SVM | Tìm siêu phẳng tối ưu phân tách dữ liệu, hiệu quả với dữ liệu chiều cao. | 86.5% | - Cân bằng giữa các chỉ số đánh giá. - F1-score ổn định. | - Huấn luyện lâu với tập lớn. - Nhạy cảm với lựa chọn tham số. |
| Random Forest | Tổ hợp nhiều cây quyết định, xử lý tốt quan hệ phi tuyến. | 85.8% | - Ổn định, ít overfitting. - Recall cao với nhãn tiêu cực. | - Sai nhãn tích cực khá nhiều. - Mô hình phức tạp hơn, khó giải thích. |
| Naive Bayes | Mô hình xác suất đơn giản, giả định đặc trưng độc lập nhau. | 85.6% | - Huấn luyện nhanh, dễ triển khai. - Precision cao với nhãn tích cực. | - Giả định không thực tế với dữ liệu text. - Nhầm lẫn nhiều ở nhãn tiêu cực. |

1. **Kết luận**

Từ kết quả so sánh, nhóm nhận thấy **Logistic Regression** cho kết quả tốt nhất với độ chính xác cao (88.4%) và cân bằng giữa các chỉ số đánh giá. Tuy vậy, các mô hình khác như **SVM**, **Random Forest** và **Naive Bayes** cũng cho hiệu quả tương đối tốt, mỗi mô hình đều có ưu điểm riêng phù hợp với từng loại dữ liệu và mục tiêu cụ thể.

### Hướng phát triển:

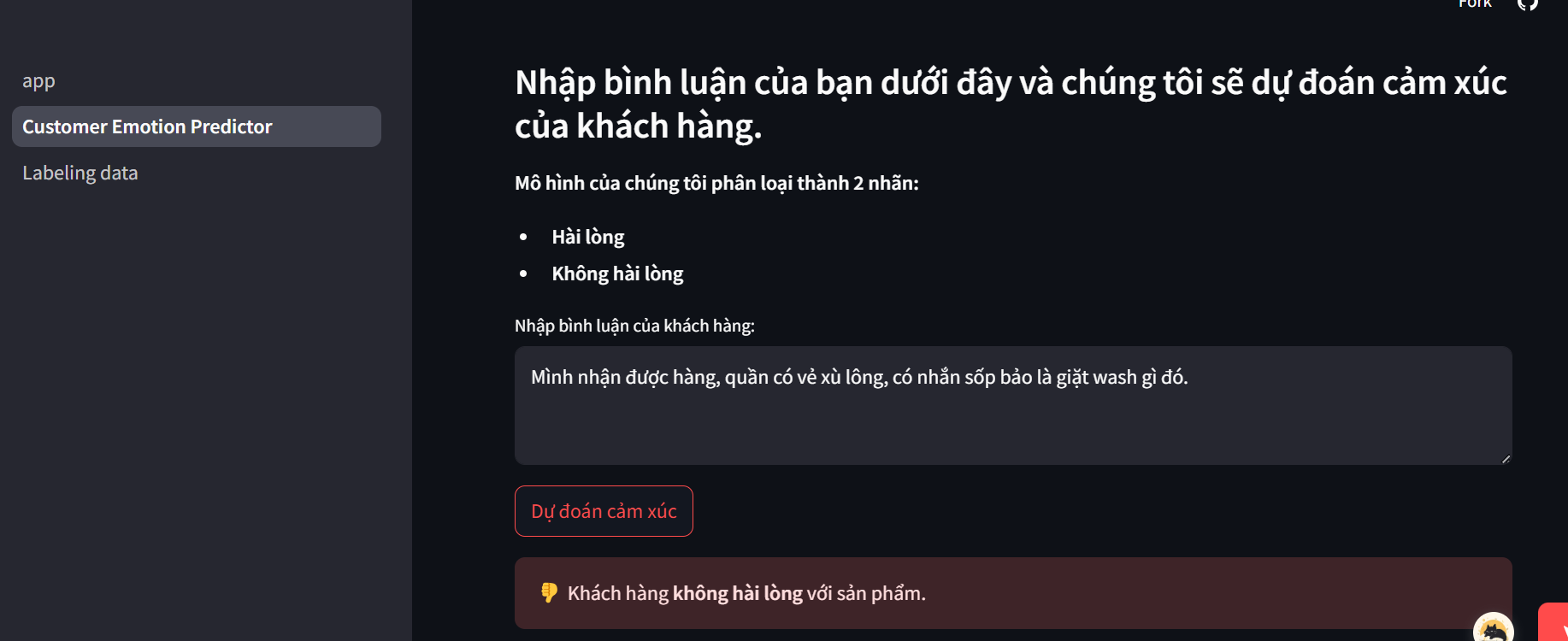
Trong tương lai, nhóm có thể thử áp dụng các mô hình **deep learning** như **LSTM**, **BERT** hoặc **Transformer** để xử lý ngữ cảnh tốt hơn và cải thiện độ chính xác trong việc nhận diện cảm xúc khách hàng từ văn bản.

# Ứng dụng

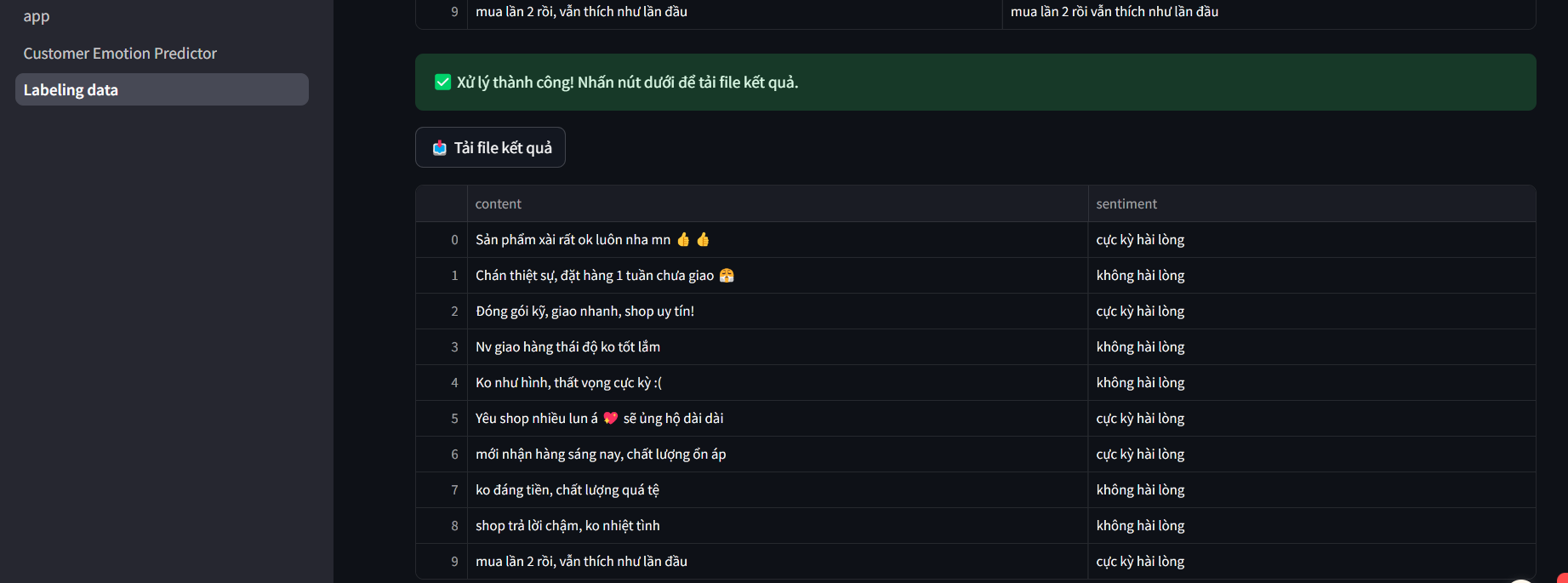
- Nhóm đã triển khai mô hình trên nền tảng Streamlit, dễ triển khai và cho phép người dùng dễ dàng tương tác và sử dụng.

- Ứng dụng hỗ trợ:

+ Dự đoán cảm xúc từ một bình luận văn bản do người dùng nhập.



+ Dự đoán cảm xúc hàng loạt từ file chỉ chứa cột content, phù hợp cho doanh nghiệp cần phân tích ý kiến khách hàng với quy mô lớn.



- Link web đã triển khai: <https://predicting-constomer-emotion.streamlit.app/>