TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CUỐI KÌ MÔN**

**CHUYÊN ĐỀ NGHIÊN CỨU KHOA HỌC DỮ LIỆU**

**PHÂN TÍCH CẢM XÚC BÀI VIẾT TRÊN FACEBOOK THEO MÔ HÌNH SENTIMENT ANALYSIS**

*Người hướng dẫn*: **PGS. TS. LÊ ANH CƯỜNG**

*Người thực hiện*: **VÕ HỮU ĐỨC – 51800536**

**NGUYỄN HỒ ĐĂNG KHOA – 51800563**

Lớp**: 18050303**

Khoá**: 22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2022**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CUỐI KÌ MÔN**

**CHUYÊN ĐỀ NGHIÊN CỨU KHOA HỌC DỮ LIỆU**

**PHÂN TÍCH CẢM XÚC BÀI VIẾT TRÊN FACEBOOK THEO MÔ HÌNH SENTIMENT ANALYSIS**

*Người hướng dẫn*: **PGS. TS. LÊ ANH CƯỜNG**

*Người thực hiện*: **VÕ HỮU ĐỨC – 51800536**

**NGUYỄN HỒ ĐĂNG KHOA – 51800563**

Lớp**: 18050303**

Khoá**: 22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2022**

LỜI CẢM ƠN

Sau hai buổi học vừa qua là những ngày mà chúng em tiếp thu rất nhiều kiến thức của thầy đã truyền đạt cho chúng em. Chúng em đã tiếp thu được những bài giảng quý báu của thầy và biết thêm nhiều kiến thức chuyên môn có thể áp dụng thưc tế cũng như đồ án của chúng em.

Để hoàn thành bài đồ án này trước tiên chúng em xin gửi đến các quý thầy, cô giảng viên trường Đại học Tôn Đức Thắng lời cảm ơn chân thành và sâu sắc nhất. Đặc biệt, em xin gửi đến thầy Lê Anh Cường – người đã tận tình hướng dẫn, giúp đỡ em hoàn thành đồ án cuối kì này lời cảm ơn sâu sắc nhất.

Với điều kiện thời gian cũng như kinh nghiệm còn hạn chế của một sinh viên, bài báo cáo này không thể tránh được những thiếu sót. Chúng em rất mong nhận được sự chỉ bảo, đóng góp ý kiến của các quý thầy cô để em có điều kiện bổ sung, nâng cao ý thức của mình, phục vụ tốt hơn công tác thực tế sau này.

Chúng em xin chân thành cảm ơn!

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là sản phẩm đồ án của riêng chúng tôi và được sự hướng dẫn của PGS.TS. Lê Anh Cường. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày tháng năm*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Võ Hữu Đức*

*Nguyễn Hồ Đăng Khoa*

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

**Phần xác nhận của GV hướng dẫn**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**Phần đánh giá của GV chấm bài**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

TÓM TẮT

Phân tích cảm xúc là cách tiếp cận được sử dụng phổ biến để phân tích dữ liệu ở dạng văn bản và xác định nội dung tình cảm từ văn bản. Các dữ liệu văn bản có thể là phản hồi, nhận xét, …. Phân tích cảm xúc là bài toán để xác định thông tin tích cực, tiêu cực, trung lập từ các đánh giá của khách hàng.

Việc phân tích dữ liệu này sẽ giúp cho bài viết có thể phân tích xem bài viết này có đáng đọc hay không từ những lời bình luận tích cực, tiêu cực hay bình thường, việc này sẽ quyết định số lương lượt đọc của bài viết đó. Từ đó dựa vào lượt đọc, bài viết đó sẽ được đưa vào dạng ưu tiên hay không ưu tiên ở trên mạng.

Thu thập dữ liệu chính xác hơn là thu thập những comments từ cư dân mạng trên facebook và sử dụng dữ liệu đã crawl được cùng với dữ liệu đã được tải về từ trên mạng đã có gán nhãn sẵn để xây dựng bộ dữ liệu sau đó tiến hành tiền xử lý dữ liệu, vector hóa dữ liệu sử dụng TF-IDF của thư viện sklearn. Chia dữ liệu để train và test, xây dựng, huấn luyện và đánh giá mô hình, sử dụng Logistic Regression của thư viện sklearn và mô hình LinearSVC. Dữ liệu sau khi xử lý trước khi đem vào để huấn luyện mô hình bao gồm 109829comments.

Xây dựng trang web bằng thư viện streamlit, được viết bằng ngôn ngữ lập trình Python.

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc108731741)

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN iii](#_Toc108731742)

[TÓM TẮT iv](#_Toc108731743)

[MỤC LỤC 5](#_Toc108731744)

[DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ 6](#_Toc108731745)

[CHƯƠNG 1 – MỞ ĐẦU 8](#_Toc108731746)

[1.1 Mô tả bài toán 8](#_Toc108731747)

[1.2 Ý nghĩa 8](#_Toc108731748)

[CHƯƠNG 2 – CÁC NGHIÊN CỨU LIÊN QUAN 9](#_Toc108731749)

[CHƯƠNG 3 – ĐỀ XUẤT GIẢI PHÁP 11](#_Toc108731750)

[3.1 Thuật toán 11](#_Toc108731751)

[3.1.1 Logistic Regression 11](#_Toc108731752)

[3.1.2 LinearSVC (Support Vector Machine) 12](#_Toc108731753)

[3.2 Kiến trúc mô hình 12](#_Toc108731754)

[3.3 Phương pháp 13](#_Toc108731755)

[CHƯƠNG 4 – THỰC NGHIỆM 16](#_Toc108731756)

[4.1 Xây dựng dữ liệu 16](#_Toc108731757)

[4.2 Kết quả thử nghiệm 19](#_Toc108731758)

[4.3 Hiện thực bài toán 20](#_Toc108731759)

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

**DANH MỤC HÌNH**

[Hình 2.1 Độ chính xác của 3 công cụ và các phương pháp học máy khác 9](#_Toc108731760)

[Hình 3.1 Decision Boundary phân loại hai dữ liệu 11](#_Toc108731761)

[Hình 3.2 Các đường siêu phẳng (Hyper-lane) chia hai lớp dữ liệu 12](#_Toc108731762)

[Hình 3.3 Kiến trúc mô hình xây dựng bài toán Sentiment Analysis 13](#_Toc108731763)

[Hình 4.1 Kết quả crawl 80000 comments từ 400 bài viết trên fanpage Theanh28 Entertainment 16](#_Toc108731838)

[Hình 4.2 Tiền xử lý và gán nhãn comments đã crawl 17](#_Toc108731839)

[Hình 4.3 Tiền xử lý file train 17](#_Toc108731840)

[Hình 4.4 Tiền xử lý file test 18](#_Toc108731841)

[Hình 4.5 Tổng hợp thành 1 file comments đã qua xử lý 18](#_Toc108731842)

[Hình 4.6 Độ chính xác mô hình LinearSVC 19](#_Toc108731843)

[Hình 4.7 Độ chính xác mô hình Logistic Regression 19](#_Toc108731844)

[Hình 4.8 Kết quả thử nghiệm 20](#_Toc108731845)

[Hình 4.9 Mở Anaconda 20](#_Toc108731846)

[Hình 4.10 Thực hiện mở terminal và tiến hành mở jupyter lab 21](#_Toc108731847)

[Hình 4.11 Giao diện sau khi mở jupyter lab 21](#_Toc108731848)

[Hình 4. 12 Trỏ đến thư mục sentiment 22](#_Toc108731849)

[Hình 4.13 Dùng câu lệnh để khởi động web 22](#_Toc108731850)

[Hình 4.14 Giao diện web phân tích cảm xúc từ bài viết Facebook 23](#_Toc108731851)

[Hình 4.15 Nhập link vào ô input 23](#_Toc108731852)

[Hình 4.16 Giao diện các bình luận đã được crawl 24](#_Toc108731853)

[Hình 4.17 Giao diện các bình luận đã được gán nhãn qua 2 mô hình 24](#_Toc108731854)

[Hình 4.18 Giao diện thống kê bình luận qua các nhãn 25](#_Toc108731855)

[Hình 4. 19 Giao diện của biểu đồ thống kê các nhãn 25](#_Toc108731856)

CHƯƠNG 1 – MỞ ĐẦU

1.1 Mô tả bài toán

Phân tích cảm xúc là cách tiếp cận được sử dụng phổ biến để phân tích dữ liệu ở dạng văn bản và xác định nội dung tình cảm từ văn bản. Các dữ liệu văn bản có thể là phản hồi, nhận xét, …. Phân tích cảm xúc là bài toán để xác định thông tin tích cực, tiêu cực, trung lập từ các đánh giá của khách hàng.

Đối với các bài viết, mỗi bài viết sẽ là về một chủ đề cụ thể như loài báo cheetah có thể đạt đến tốc độ cực đại là 112 – 120km/h không? , …. Cũng như các cổng thương mại, mỗi bài viết sẽ đem lại nhiều câu bình luận từ những độc giả khác nhau. Việc phân tích dữ liệu này sẽ giúp cho bài viết có thể phân tích xem bài viết này có đáng đọc hay không từ những lời bình luận tích cực, tiêu cực hay bình thường, việc này sẽ quyết định số lương lượt đọc của bài viết đó. Từ đó dựa vào lượt đọc, bài viết đó sẽ được đưa vào dạng ưu tiên hay không ưu tiên ở trên mạng.

1.2 Ý nghĩa

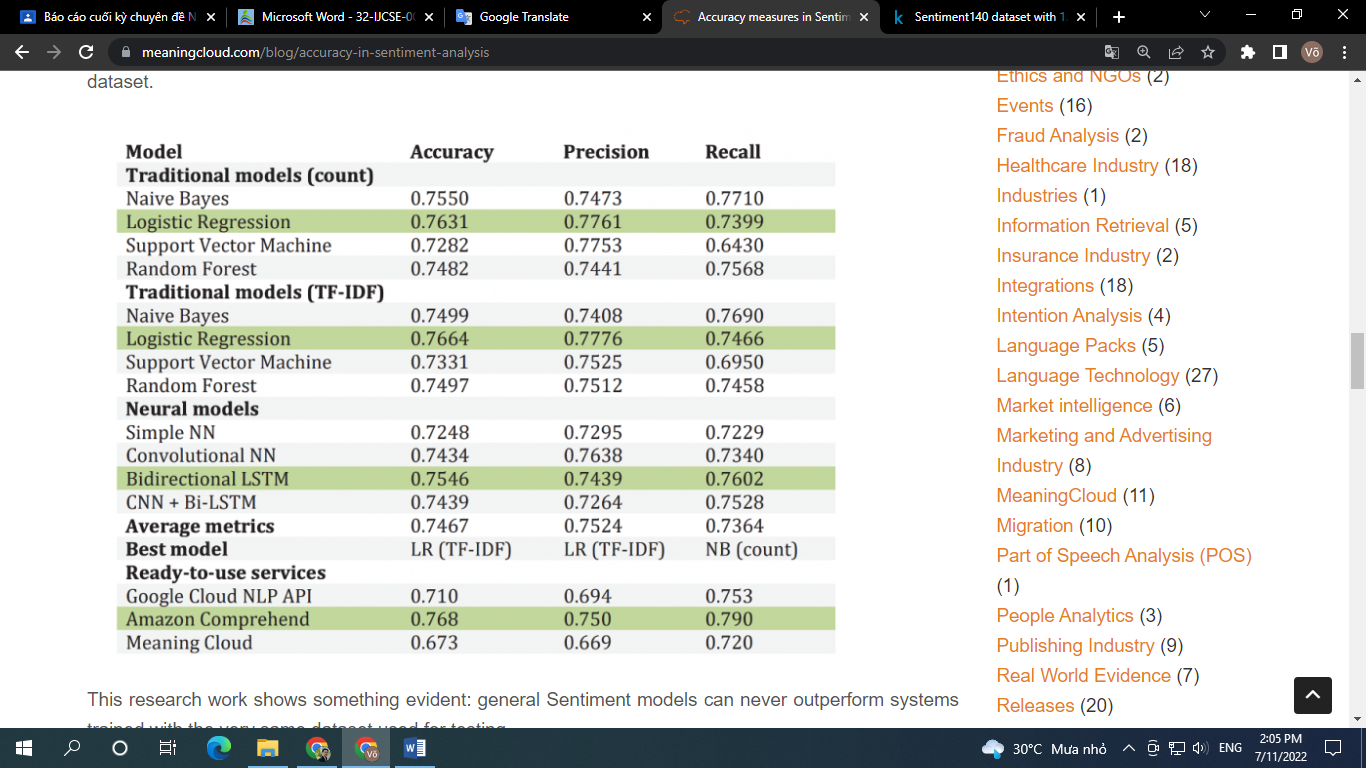
Việc phân tích cảm xúc từ các bình luận về một bài viết sẽ giúp cho các độc giả khác có thể quyết định một cách cẩn thận khi muốn đọc bài viết đó. Nếu như những bình luận của bài viết chiếm tỷ lệ tiêu cực nhiều thì cư dân mạng đó sẽ tránh đọc những bài viết đó và ngược lại những bài viết mang xu hướng tích cực sẽ được nhiều cư dân mạng biết đến và đọc nhiều hơn.

Chính vì thể nên nhóm chúng em chọn đề tài “Phân tích và dự đoán cảm xúc bình luận về một bài viết trên Facebook theo mô hình Sentiment Analysis”.

CHƯƠNG 2 – CÁC NGHIÊN CỨU LIÊN QUAN

Trong một bài luận văn của Paolo Romeo, anh đã so sánh ba công cụ thương mại (Google Cloud NLP API, Amazon Comprehend và MeaningCloud) với các phương pháp truyền thống học máy khác.

Anh sử dụng cơ sở dữ liệu Sentiment140 với 1 triệu 6 dòng tweet (vừa có tích cực, vừa có tiêu cực và trung hòa, trong đó các tweet từ 15 chữ trở lên). Cơ sở dữ liệu Sentiment140 là cơ sở dữ liệu phân tích tình cảm được gắn thẻ lớn nhất, là nguồn tham khảo được sử dụng rộng rãi để đào tạo.



Hình 2.1 Độ chính xác của 3 công cụ và các phương pháp học máy khác

Ta có thể thấy, ngoại trừ phương pháp MeaningCloud ra thì các phương pháp học máy khác đều cho ra độ chính xác từ 0.7 trở lên. Mặc dù MeaningCloud chỉ cho kết quả là 0.673(67.3%), Paolo đánh giá rằng đây là kết quả xuất sắc cho MeaningCloud bởi vì MeaningCloud là giải pháp duy nhất trong việc so sánh khi mà tập test của nó không nằm trong việc đào tạo của hệ thống.

Từ đó ta có thể rút ra được kết luận: các mô hình Cảm xúc chung không bao giờ có thể hoạt động tốt hơn các hệ thống được đào tạo với cùng một bộ dữ liệu được sử dụng để thử nghiệm.

CHƯƠNG 3 – ĐỀ XUẤT GIẢI PHÁP

3.1 Thuật toán

Thuật toán sẽ được sử dụng trong bài toán của nhóm em là Logistic Regression và LinearSVM.

3.1.1 Logistic Regression

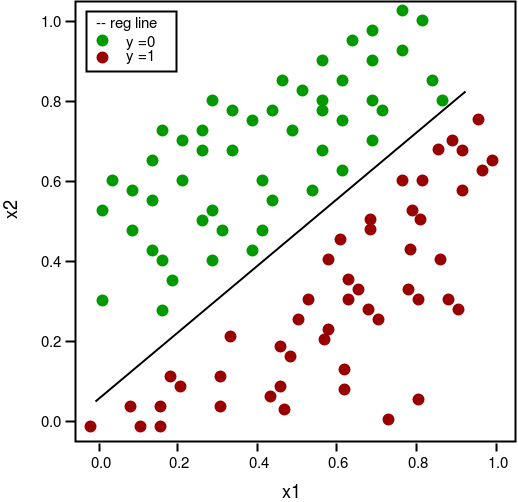
Logistic Regression là một phương pháp phân tích thống kê được sử dụng để dự đoán các giá trị dữ liệu dựa trên các quan sát trước đó của một tập dữ liệu và là một thuật toán đơn giản nhưng hiệu quả cao cho các bài toán phân loại (Classification).

Mục đích của Logistic Regression là ước tính xác suất của một sự kiện, bao gồm xác định mối quan hệ giữa các đối tượng để dự đoán xác suất kết quả.

Vì vậy, đối với Logistic Regression sẽ có:

Input: Dữ liệu input (hai nhãn 0 và 1)

Output: Xác xuất dữ liệu input rơi vào nhãn 0 hoặc là nhãn 1



Hình 3.1 Decision Boundary phân loại hai dữ liệu

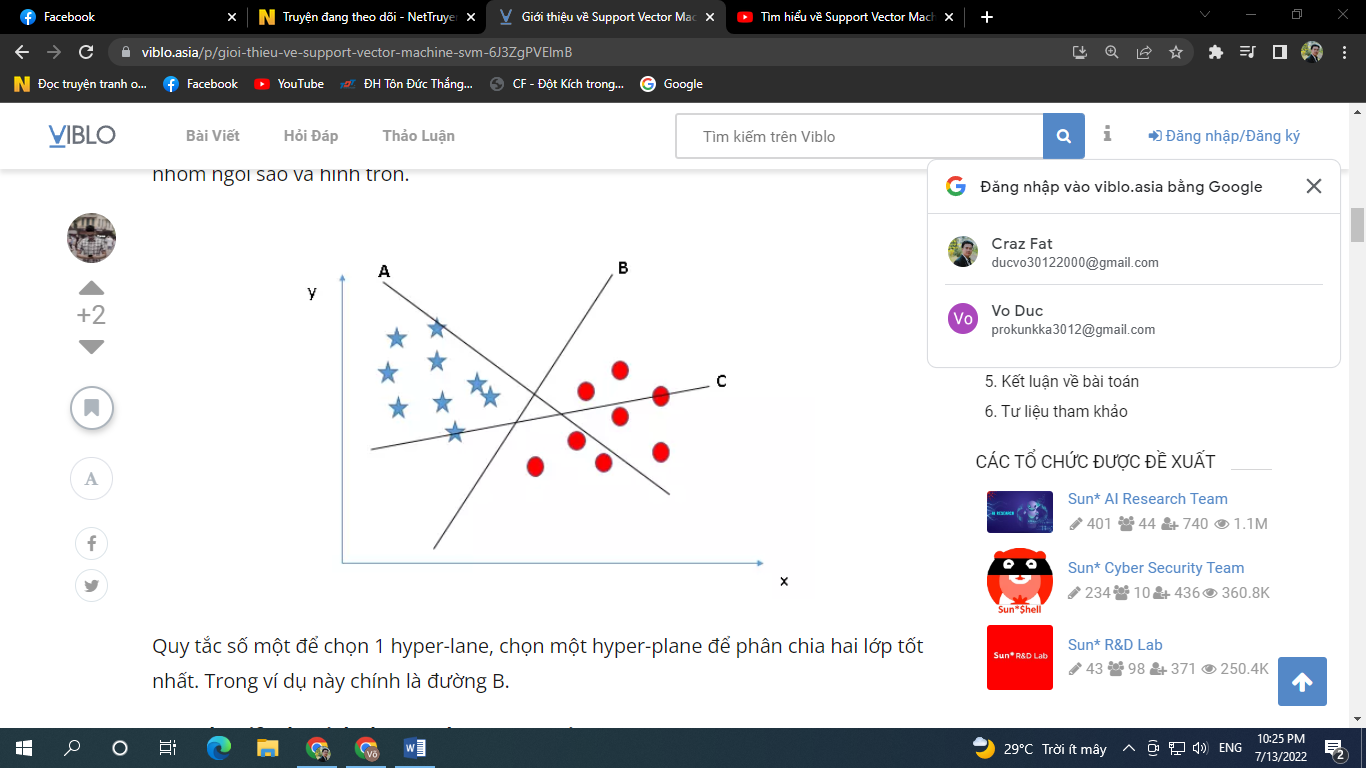
Các điểm dữ liệu màu xanh có giá trị bằng 0 và các điểm dữ liệu màu đỏ có giá trị bằng 1, ta sẽ biết được xác suất để mỗi điểm rơi trên nhãn 0 và xác suất rơi trên nhãn 1 là bao nhiêu.

3.1.2 LinearSVC (Support Vector Machine)

Linear thực chất được import từ thư viện sklearn của SVM, SVM là một thuật toán được giám sát, là phương pháp phân lớp nhanh, nó được dụng để phân chia dữ liệu thành các nhóm riêng biệt. Ứng dụng của SVM gồm: Nhận dạng giọng nói, phân loại văn bản, v.v….

SVM là thuật toán được sử dụng rất hiệu quả cho việc giải quyết các bài toán có dữ liệu lớn, rất tốt trong việc giải quyết vấn để Overfitting, có hiệu suất tổng hợp tốt và tính toán cao.

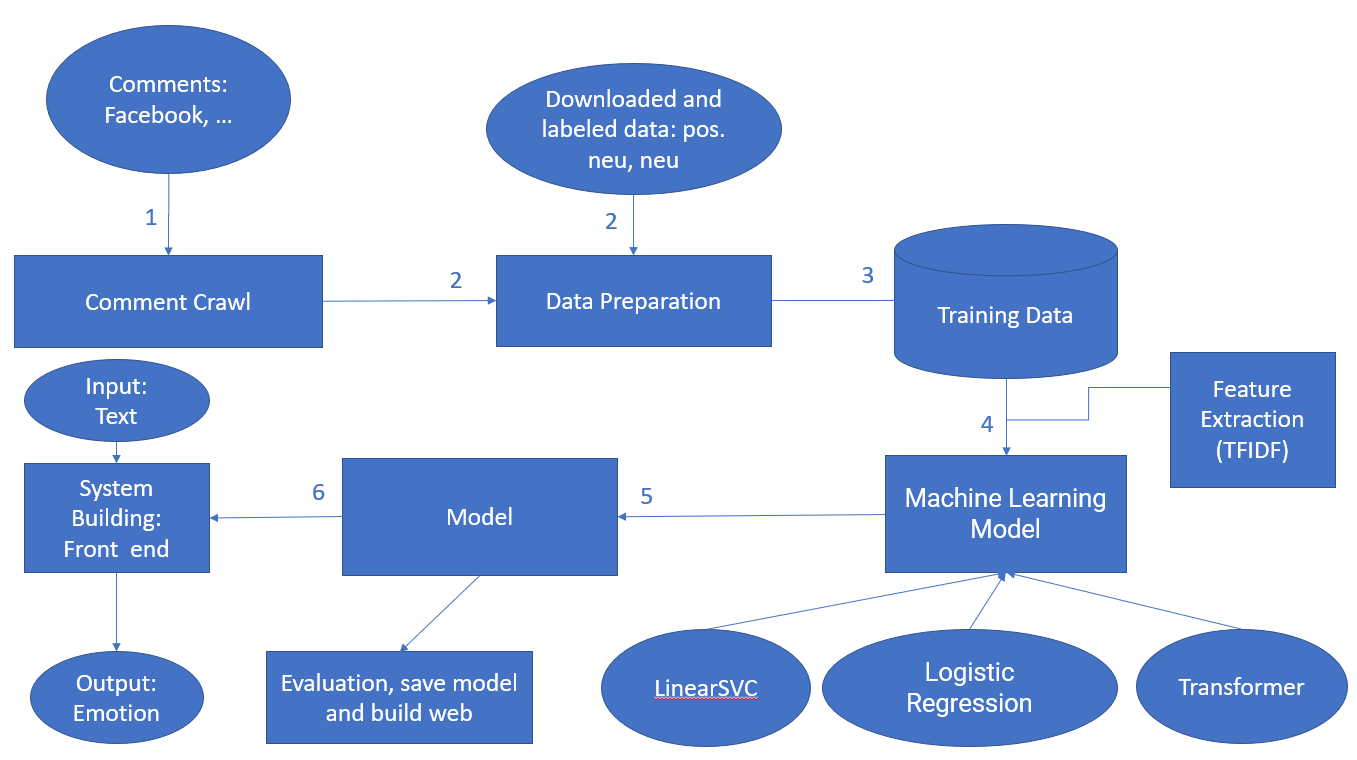
Phương pháp Support Vector Machine là tìm một siêu phẳng (hyperplane) để phân tách các điểm dữ liệu. Siêu phẳng này sẽ chia không gian thành các miển khác nhau và mỗi miền sẽ chứa một loại dữ liệu.



Hình 3.2 Các đường siêu phẳng (Hyper-lane) chia hai lớp dữ liệu

Để chọn một Hyperplane để phân chia dữ liệu hai lớp tốt nhất, ta có thể thấy đường thẳng B là lựa chọn tốt nhất để phân chia hai lớp đó.

3.2 Kiến trúc mô hình



Hình 3.3 Kiến trúc mô hình xây dựng bài toán Sentiment Analysis

3.3 Phương pháp

Theo như hình 3.1, phương pháp cho bài toán Sentiment Analysis như sau:

1. **Thu thập dữ liệu**:

* Chính xác hơn đó chính là thu thập comments từ những người dùng facebook bằng cách sử dụng **Selenium** cùng với **chrome driver** để mô phỏng trình duyệt để có thể crawl comments, sử dụng **cookie** để đăng nhập facebook. Nhóm chúng em đã crawl từ **400** bài viết từ **fanpage Theanh28 Entertainment** mỗi bài viết chúng em chỉ lấy **200** comments để có thể thu thập dữ liệu được đa dạng hơn. Mặc dù có những khó khăn về mặt thời gian do thời gian crawl quá lâu, nhưng nhóm chúng em cũng đã thu thập được chính xác **80000** comments và lưu thành file .csv để có thể tiền xử lý dữ liệu trước khi train model.
* Nói rõ hơn về Selenium, Selenium được sử dụng để automate các thao tác với trình duyệt, hay dễ hiểu hơn là nó giúp giả lập lại các tương tác trên trình duyệt như một người dùng thực sự.

1. **Tiền xử lý dữ liệu:**

* Đọc file .csv có comments crawl từ Facebook, cũng như đọc file .csv từ dữ liệu download từ trên mạng về để tiến hành tiền xử lý dữ liệu.
* Nhóm chúng em đã xử lý bằng một số cách như: xóa cột index, xóa những dòng có NaN và empty, đưa comments về chữ thường, xóa emoji, xóa vietnamese stopwords, word segmentation. Và cuối cùng lưu thành file .csv.
* Đối với comments crawl từ facebook phải gán nhãn cho từng comment.

1. **(4) Đưa dữ liệu vào để train model**

* Tiền xử lý dữ liệu xong, nhóm chúng em đã tổng hợp thành 1 file .csv chứa những comment đã được gán nhãn.
* Tiến hành đọc file .csv, chia tập train thành 20% test, 80% là train và xử lý dữ liệu để có thể đưa được vào 2 models: LinearSVC và Logistic Regression.

**(5) Độ chính xác và lưu models**

* Sử dụng 2 models để dự đoán thu được độ chính xác như sau: **LinearSVC** với độ chính xác **74%** và **Logistic Regression** với độ chính xác **75%**. Nhóm chúng em thấy Logistic Regression thu được kết quả tốt hơn.
* Sau đó tiến hành lưu 2 models bằng cách sử dụng **joblib**, bởi nếu không lưu mà tiến hành đưa code vào web thì quá trình train có lẽ sẽ mất rất nhiều thời gian.
* Nói rõ hơn **joblib** chính là một phần của hệ sinh thái SciPy, nó cũng hỗ trợ việc lưu ML model thành file rất dễ dàng, sử dụng cấu trúc dữ liệu của NumPy. Ưu điểm của việc sử dụng joblib so với pickle là nó hoạt động khá nhanh, đặc biệt với những model có kích thước lớn.

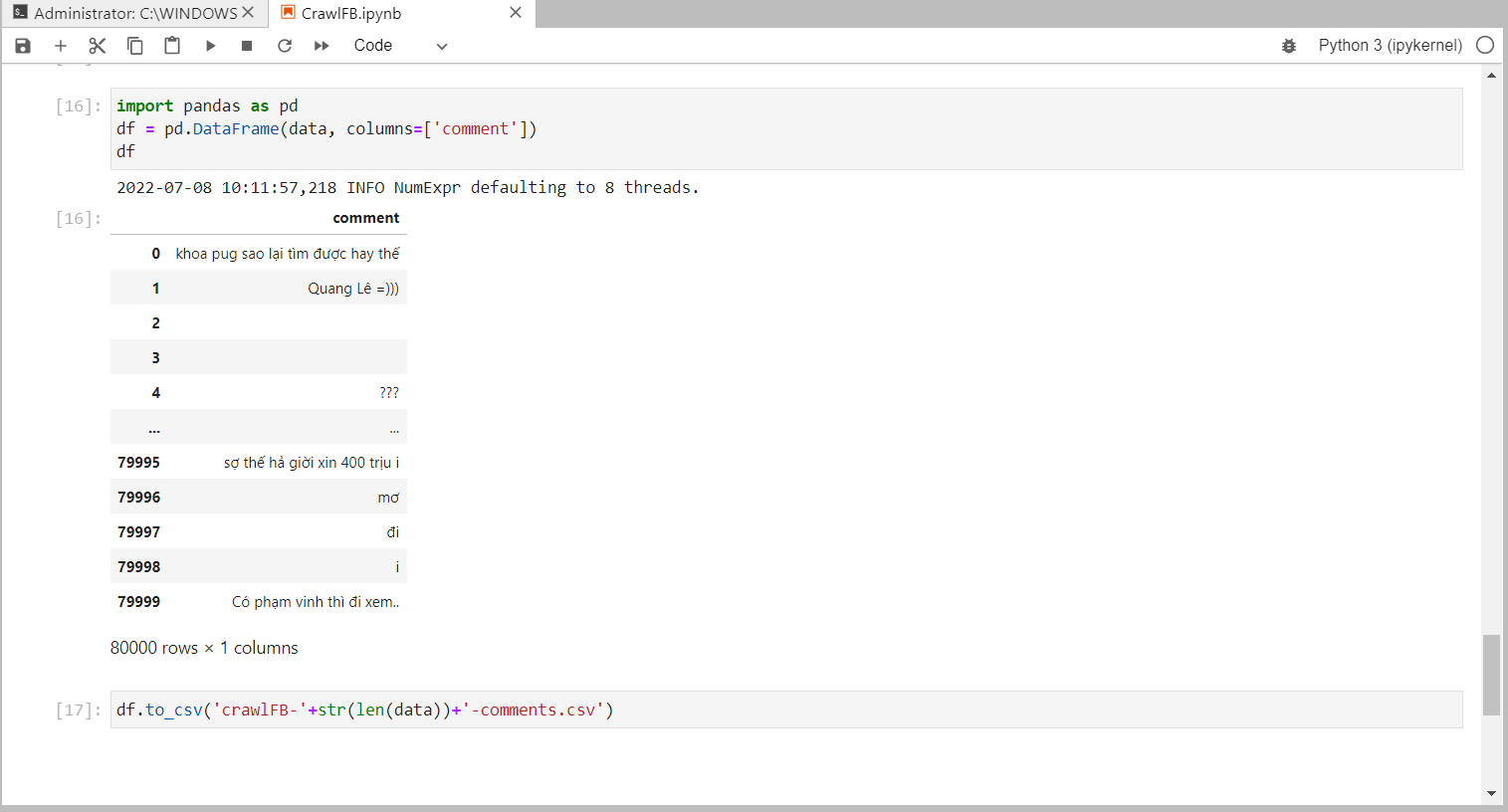
**(6) Xây dựng web**

* Nhóm chúng em đã xây dựng web bằng cách sử dụng thư viện **streamlit** của python.
* Web sẽ có 1 ô input để người dung có thể nhập vào link từ một bài viết trên Facebook. Sau khi nhấn enter thì đợi vài phút để crawl comments và được kết quả là những comment đã được gán nhãn. Những nhãn bao gồm: **pos** **(tích cực), neg (tiêu cực), neu (trung tính).** Và hiển thị lên một số tổng hợp như là có bao nhiêu comments cho từng nhãn, biểu đồ thống kê % cho 3 loại nhãn đó.

CHƯƠNG 4 – THỰC NGHIỆM

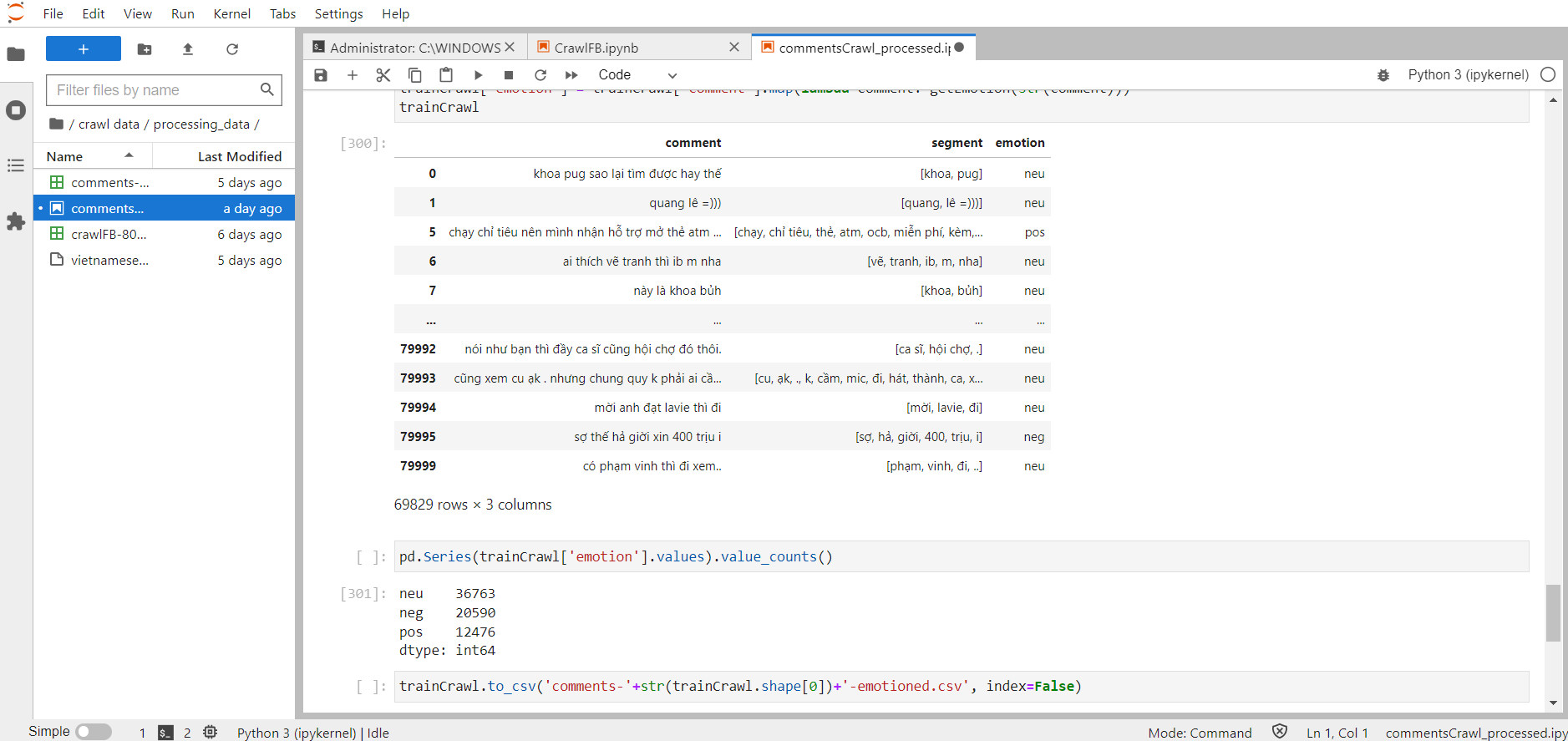
4.1 Xây dựng dữ liệu

Nhóm chúng em đã crawl từ **400** bài viết từ **fanpage Theanh28 Entertainment** mỗi bài viết chúng em chỉ lấy **200** comments để có thể thu thập dữ liệu được đa dạng hơn. Mặc dù có những khó khăn về mặt thời gian do thời gian crawl quá lâu, nhưng nhóm chúng em cũng đã thu thập được chính xác **80000** comments và lưu thành file **crawlFB-80000-comments.csv** để có thể tiền xử lý dữ liệu trước khi train model.



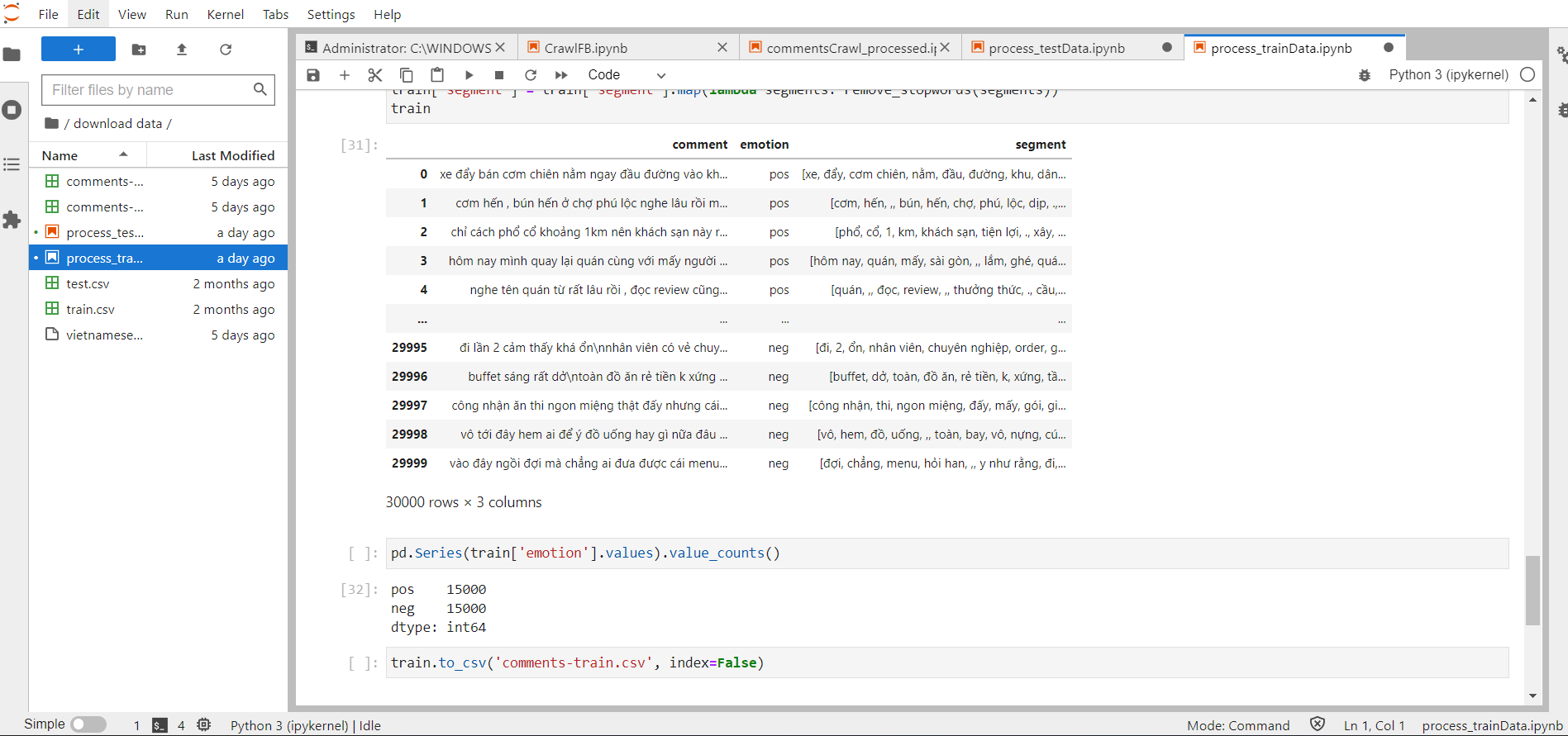
Hình 4.1 Kết quả crawl 80000 comments từ 400 bài viết trên fanpage Theanh28 Entertainment

Tiền xử lý và sau đó gán nhãn cho 80000 comments, sau khi xử lý xong chỉ còn 69829 comments.

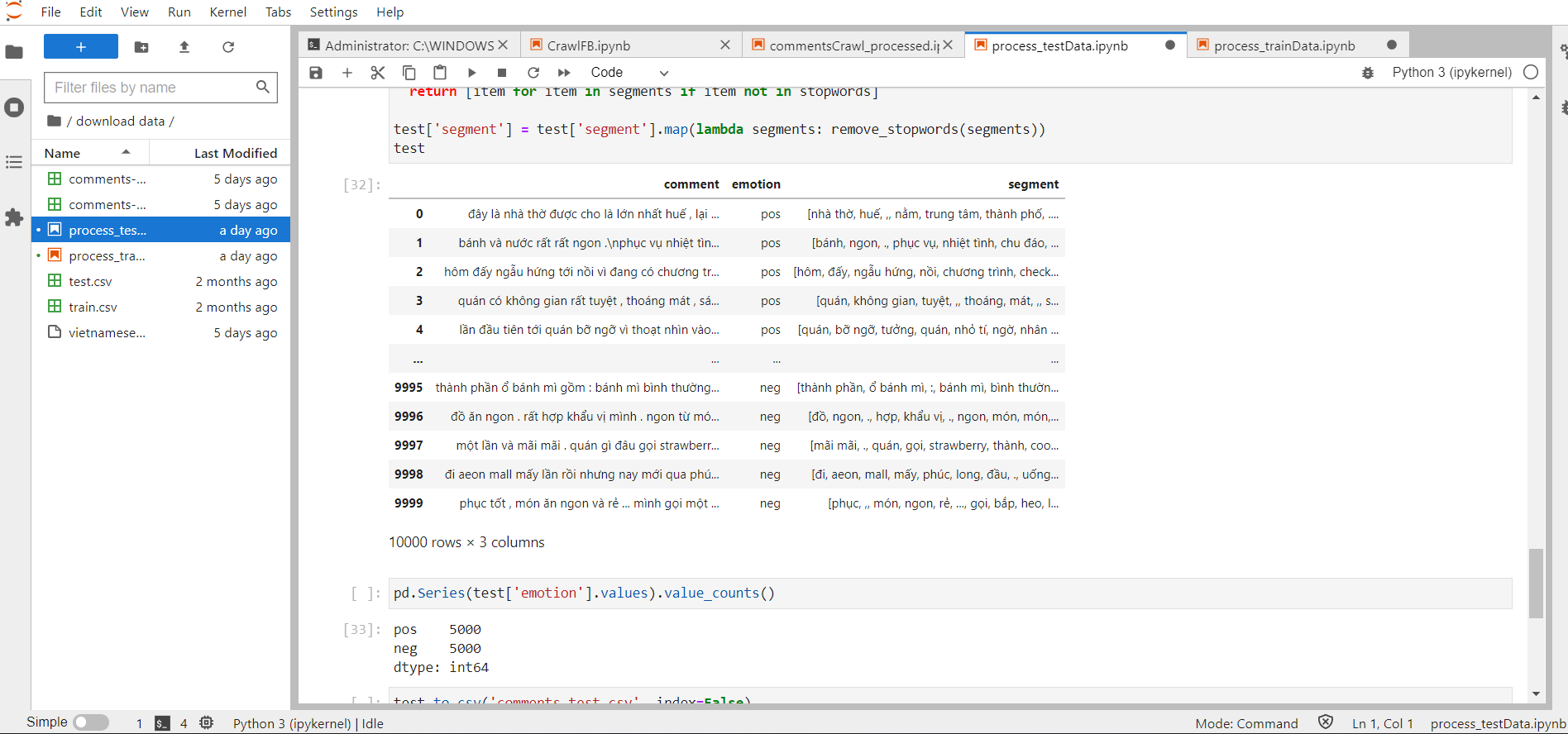


Hình 4.2 Tiền xử lý và gán nhãn comments đã crawl

Tiền xử lý cho 2 file train và test đã được tải về và đã được gán nhãn sẵn.

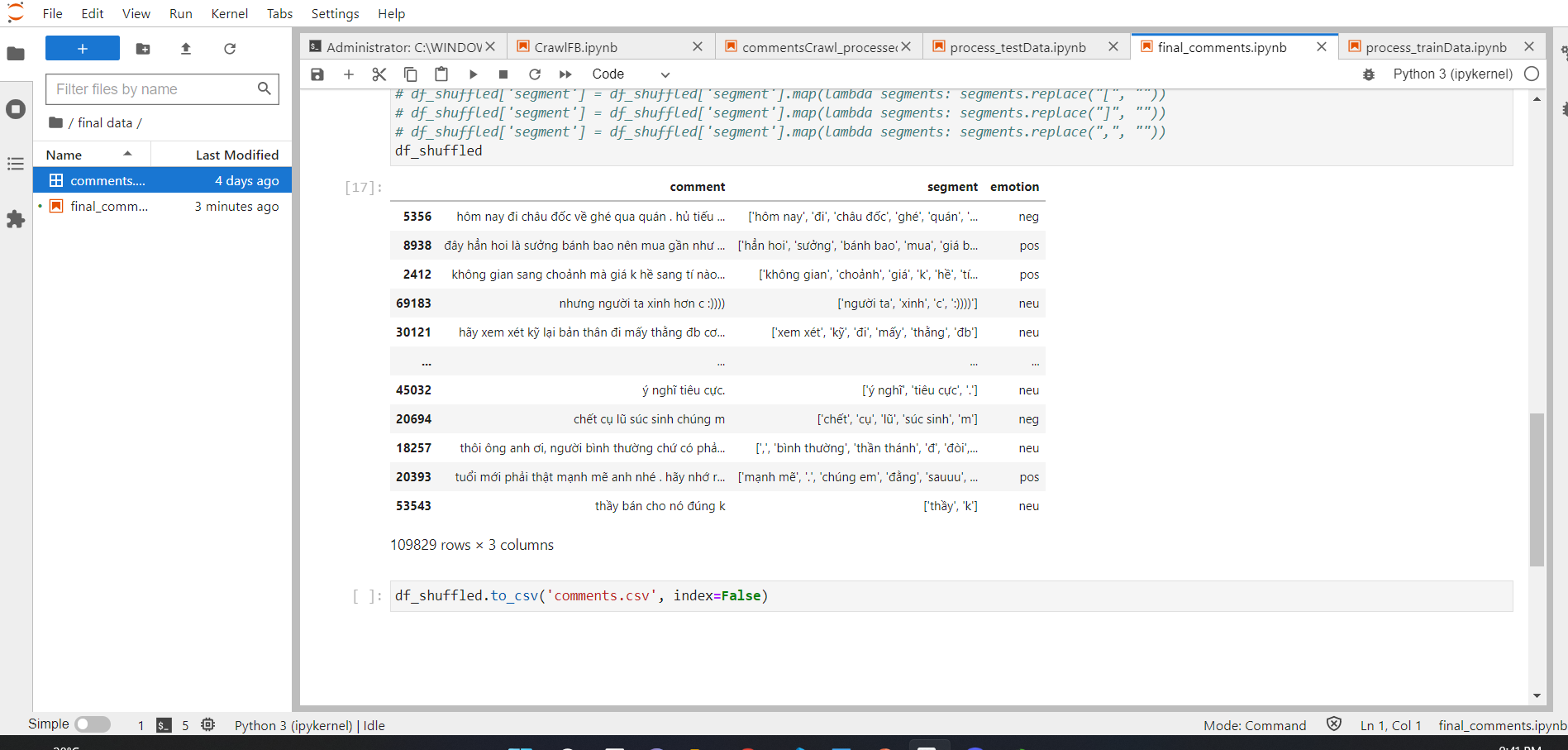


Hình 4.3 Tiền xử lý file train



Hình 4.4 Tiền xử lý file test

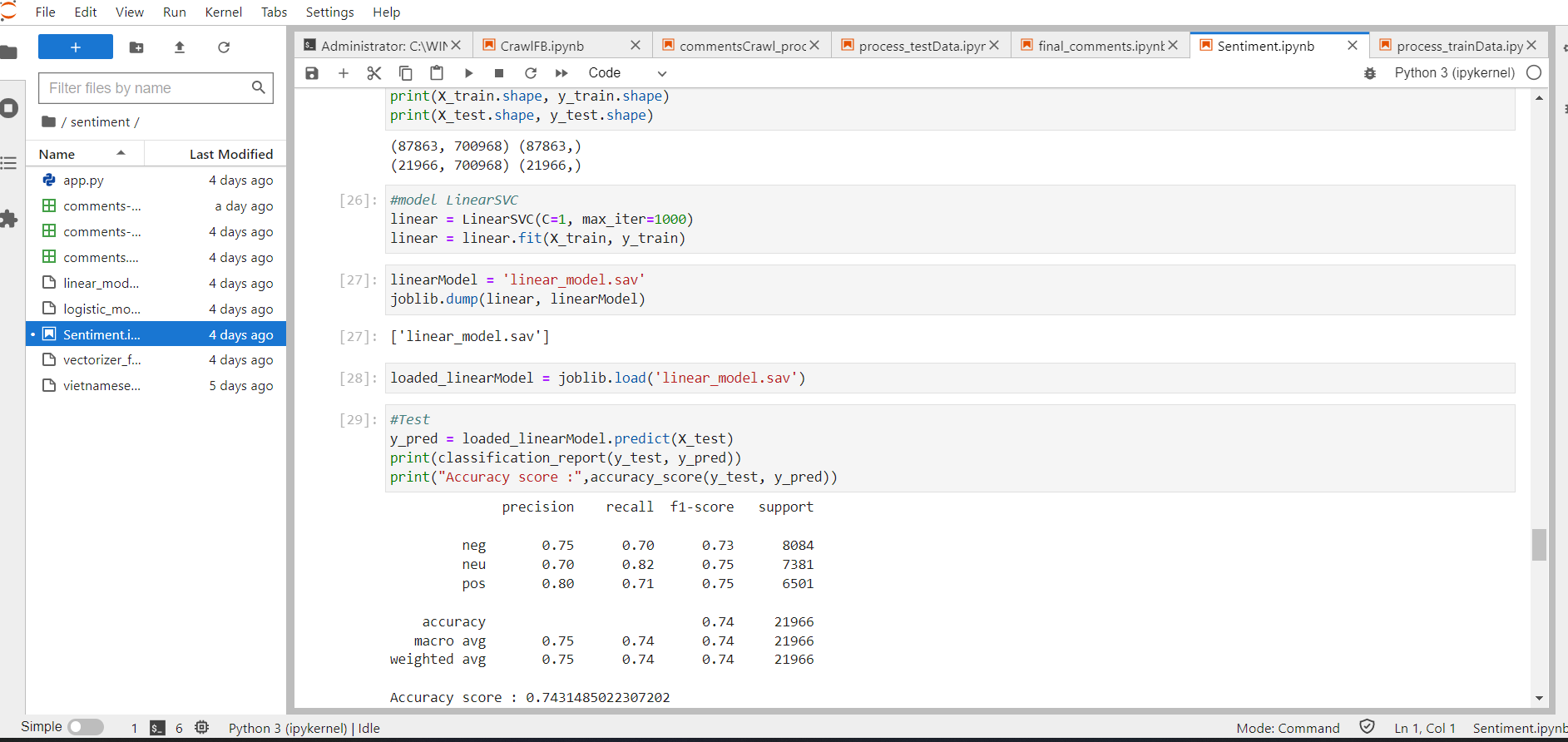
Tiến hành gộp những file trên thành 1 file **comments.csv** bằng cách đọc 3 file .csv và dùng **concat** để có thể nối chúng lại và sử dụng **shuffle** để trộn thứ tự những comment lại và kết quả nhóm chúng em nhận được file **comments.csv** có **109829** rows. Cuối cùng để có thể sử dụng file **comments.csv** để có thể train models



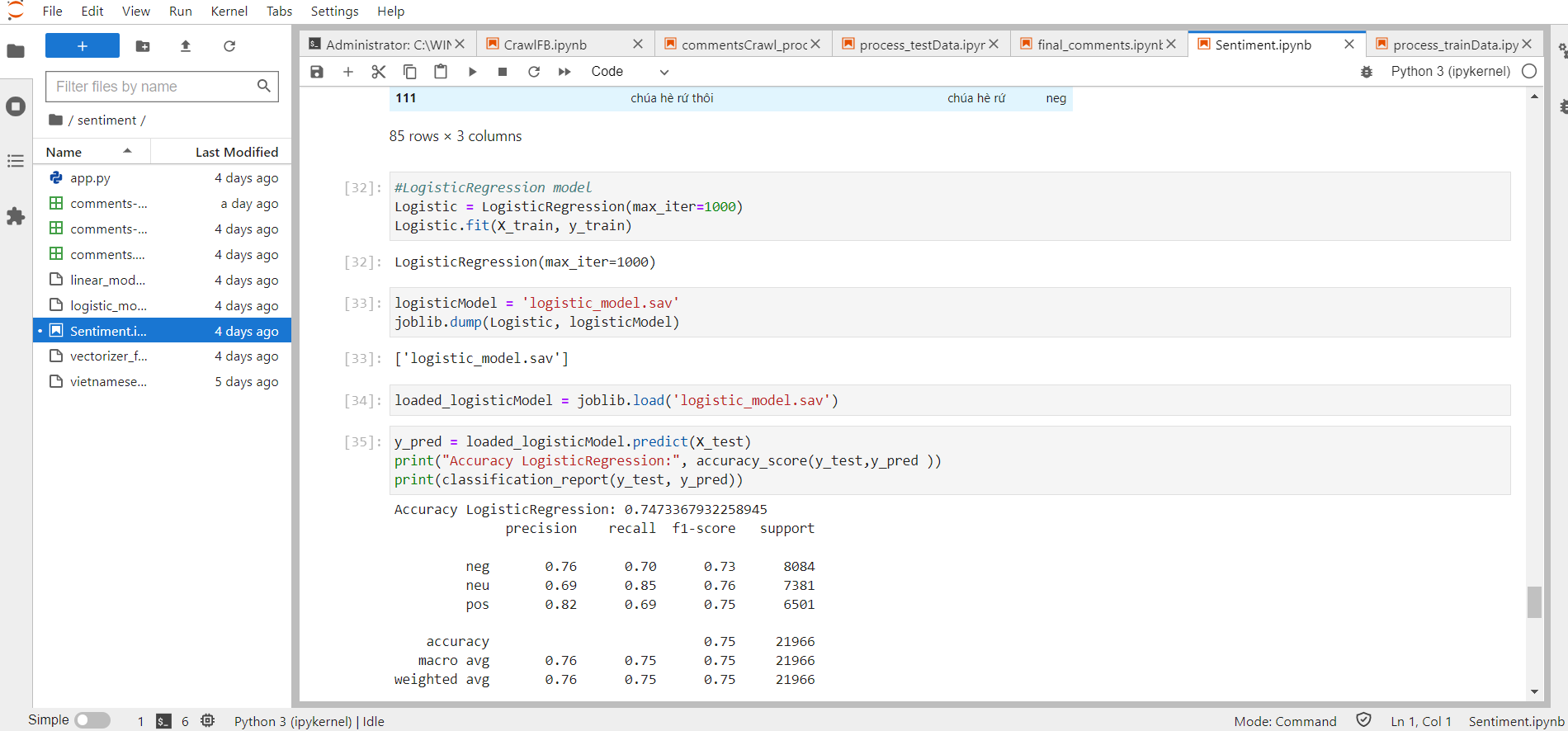
Hình 4.5 Tổng hợp thành 1 file comments đã qua xử lý

4.2 Kết quả thử nghiệm

Độ chính xác của 2 models: **LinearSVC** và **Logistic Regression**

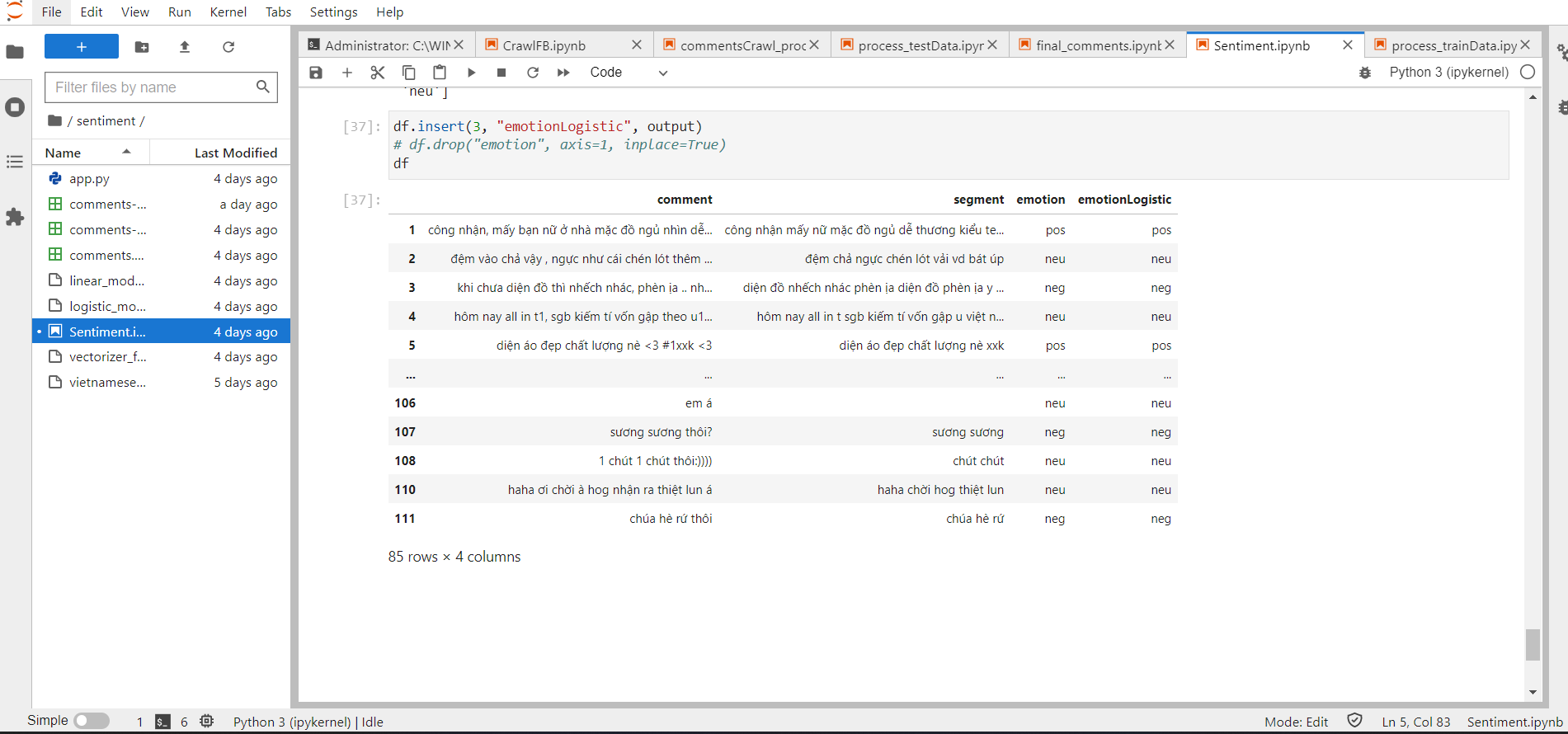


Hình 4.6 Độ chính xác mô hình LinearSVC



Hình 4.7 Độ chính xác mô hình Logistic Regression

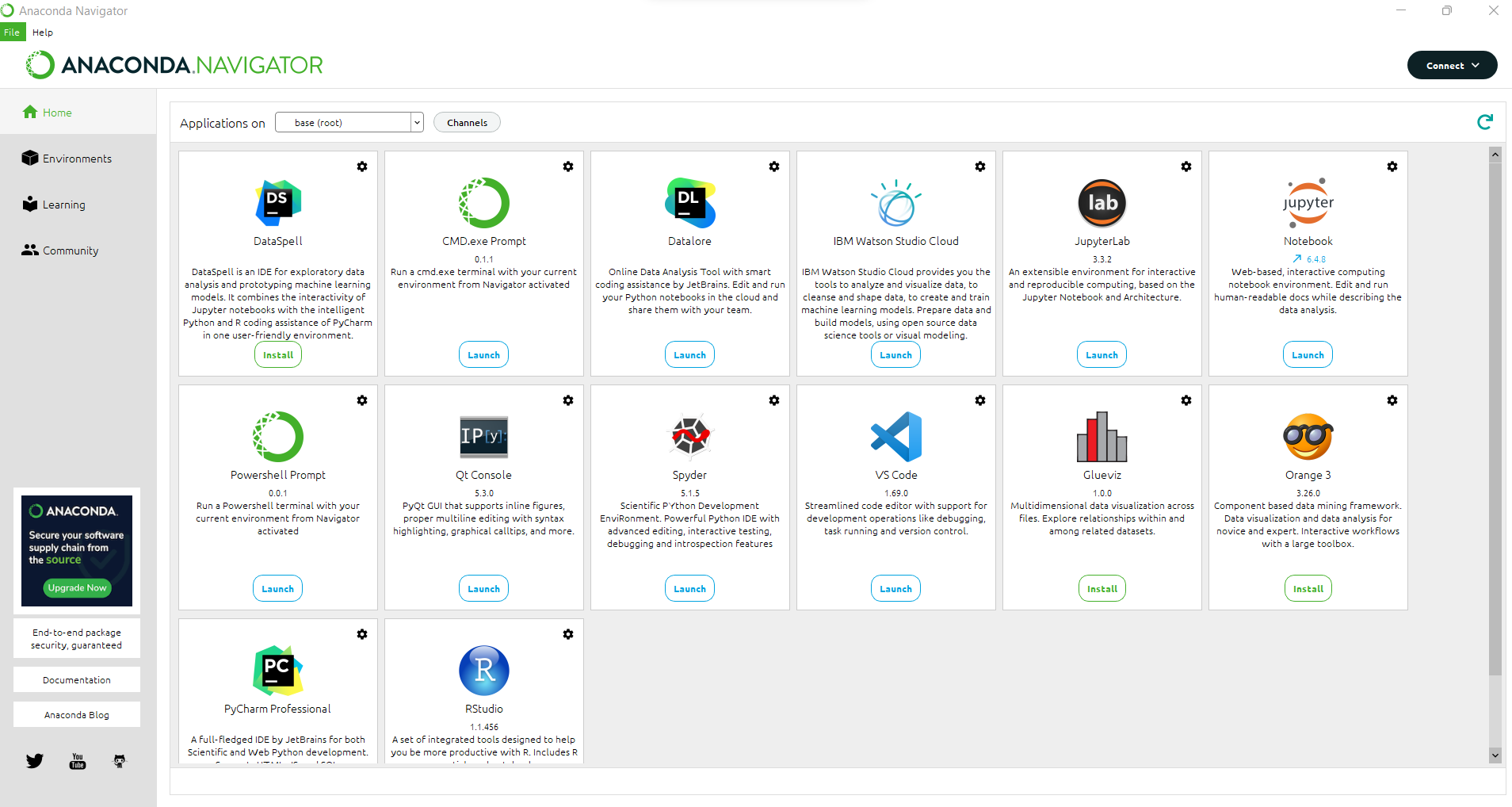
Và kết quả khi crawl thử comments từ 1 bài viết https://www.facebook.com/ Theanh28/photos/a.1509486855763896/6306054136107120 và đưa nó vào 2 mô hình.



Hình 4.8 Kết quả thử nghiệm

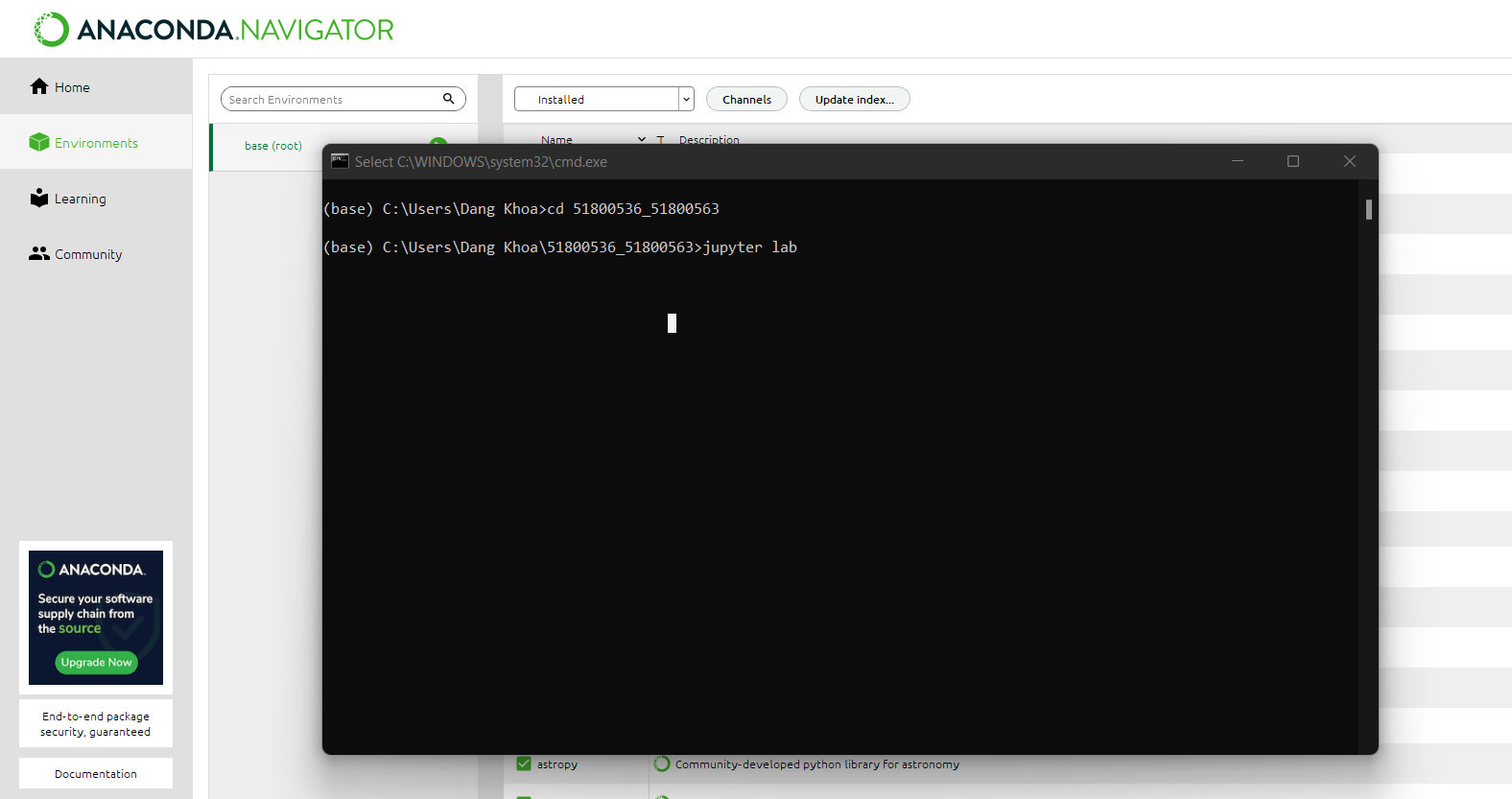
4.3 Hiện thực bài toán

Mở Anaconda và sau đó mở jupyter lab trên anaconda

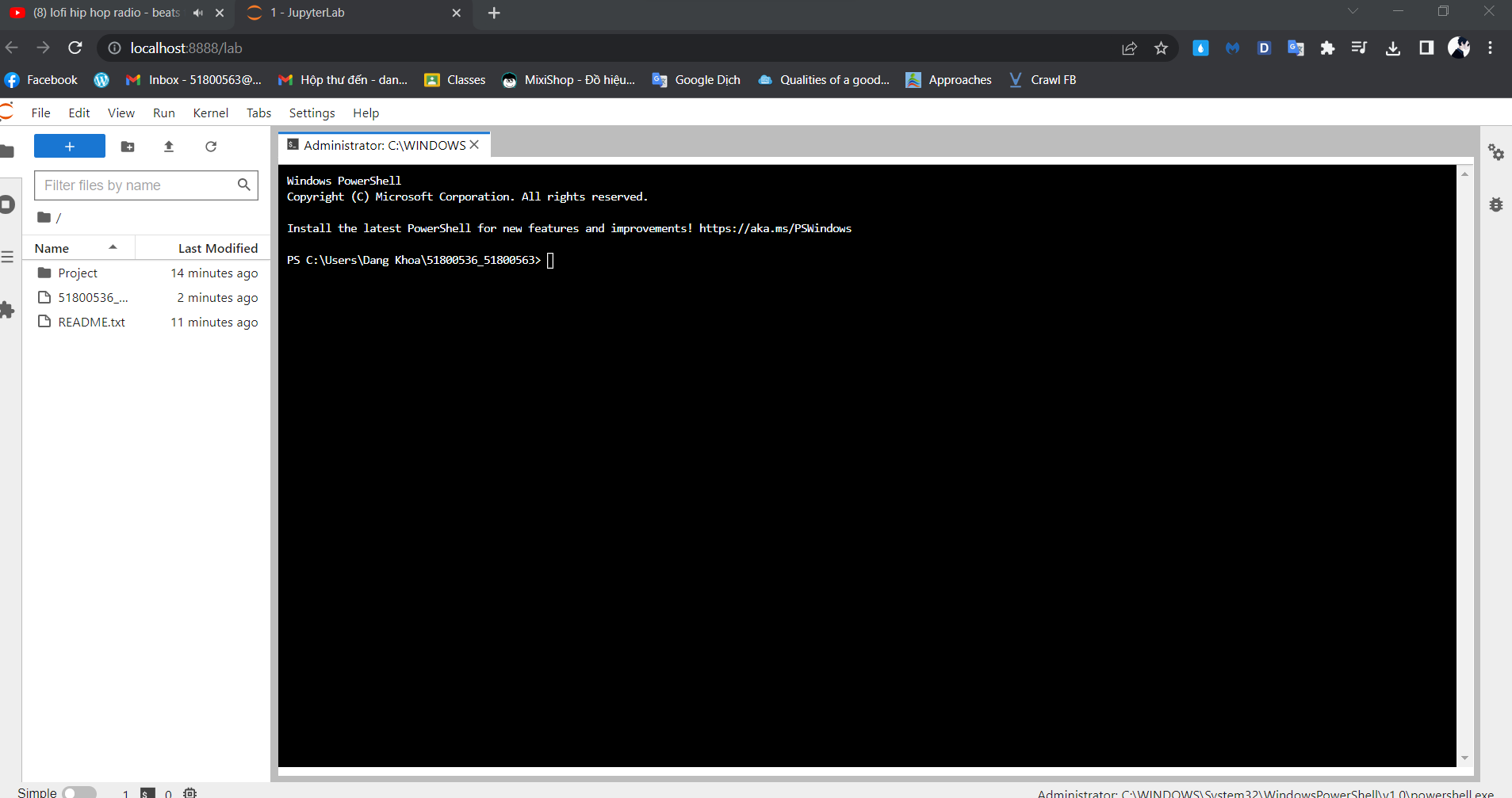


Hình 4.9 Mở Anaconda

Mở Terminal trong Anaconda và gõ dòng lệnh chuyển đến thư mục 51800536\_51800563 sau đó gõ dòng lệnh như hình dưới đây:

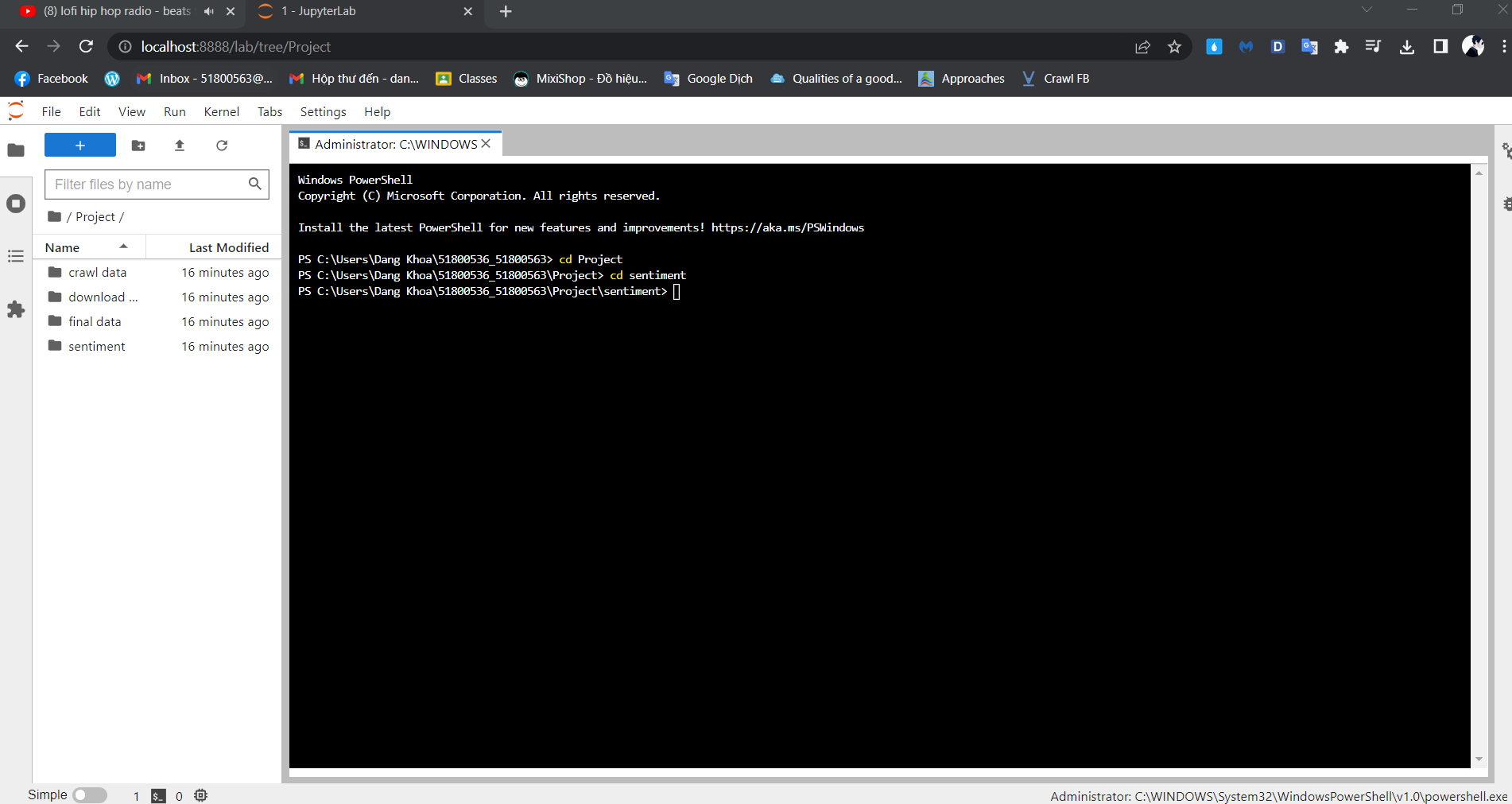


Hình 4.10 Thực hiện mở terminal và tiến hành mở jupyter lab



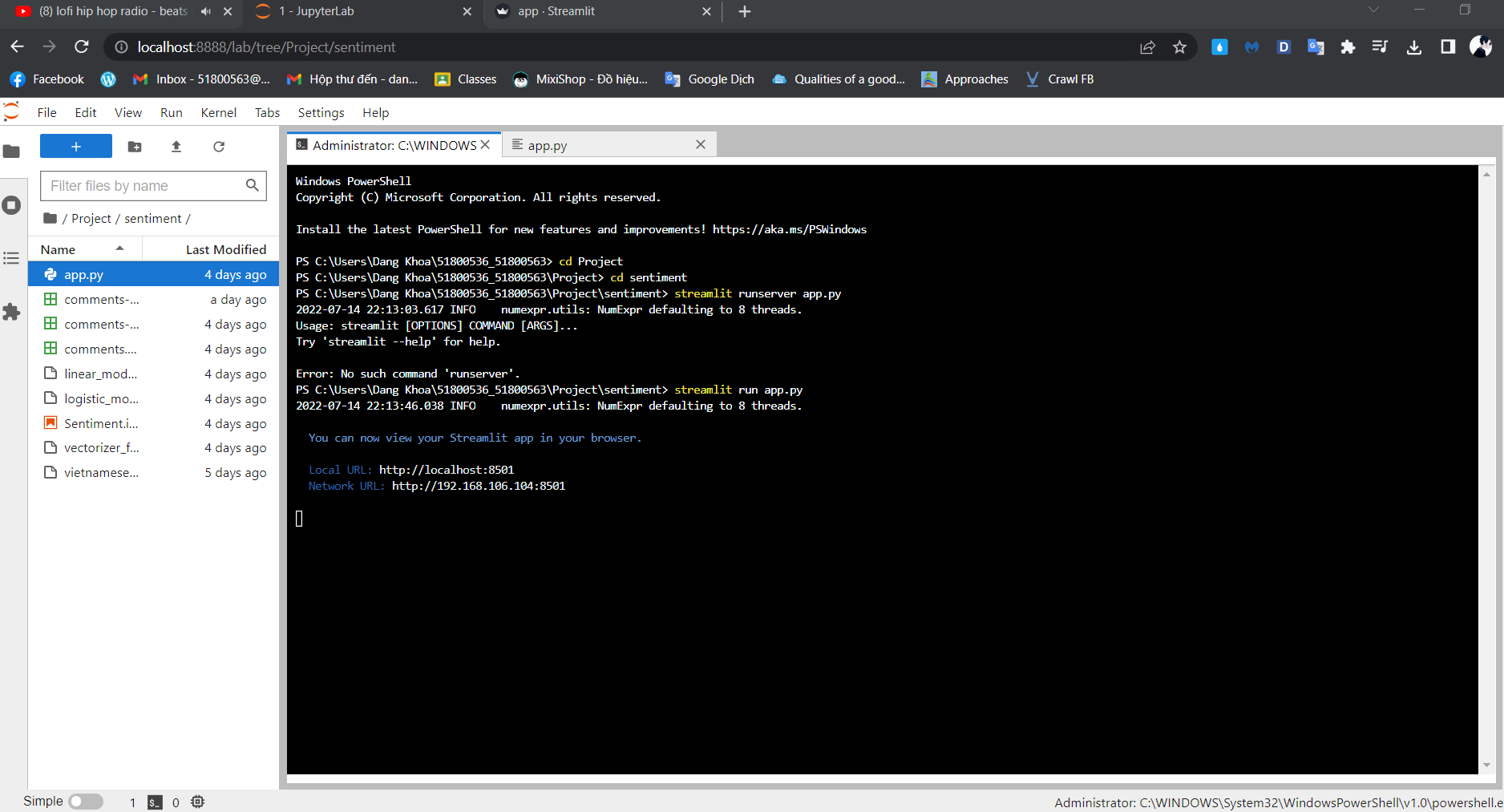
Hình 4.11 Giao diện sau khi mở jupyter lab

Sử dụng terminal của jupyter lab và trỏ đến thư mục sentiment



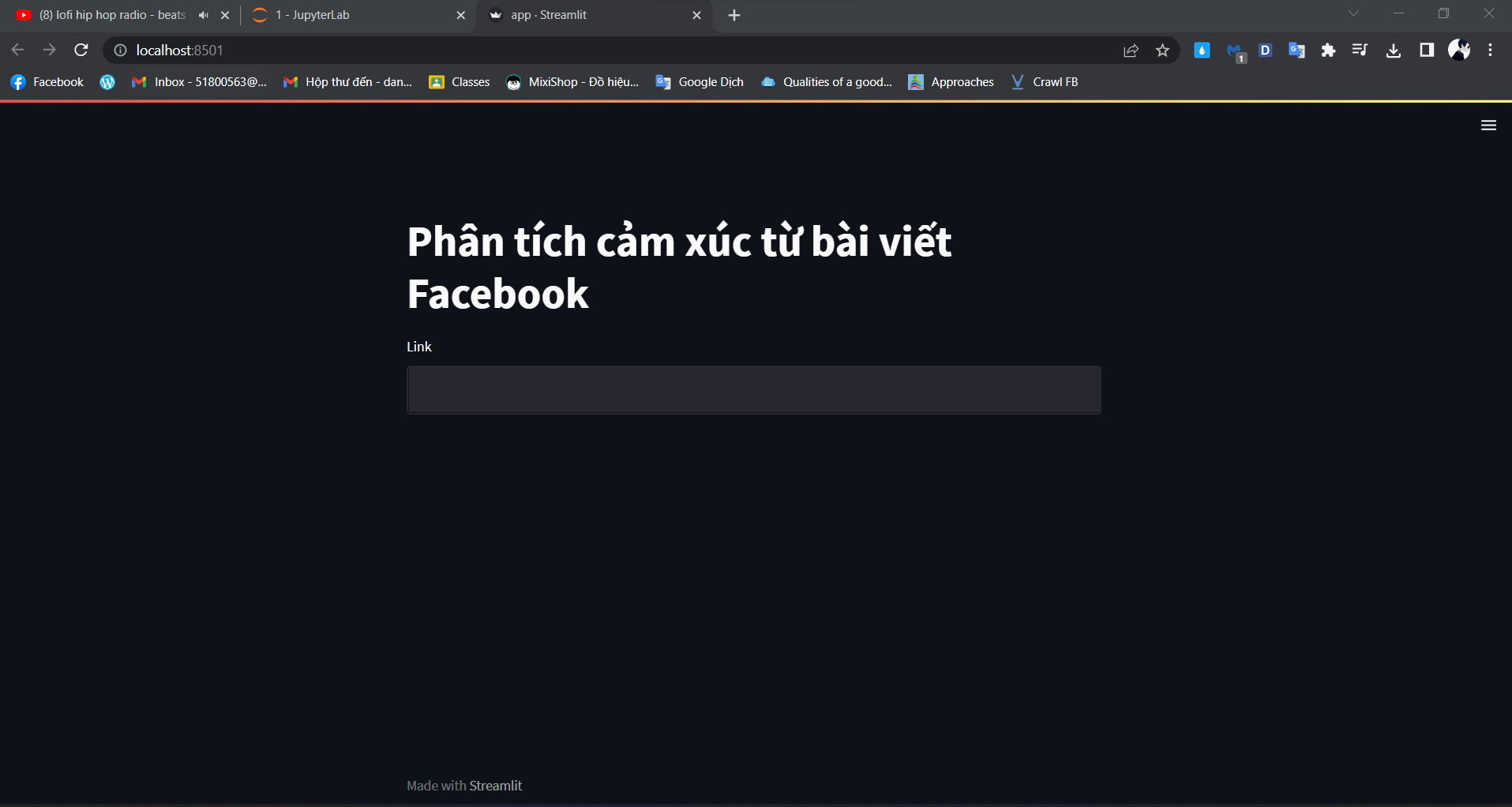
Hình 4. 12 Trỏ đến thư mục sentiment

Sau đó sử dụng dòng lệnh để chạy web, file để chạy web có tên là app.py



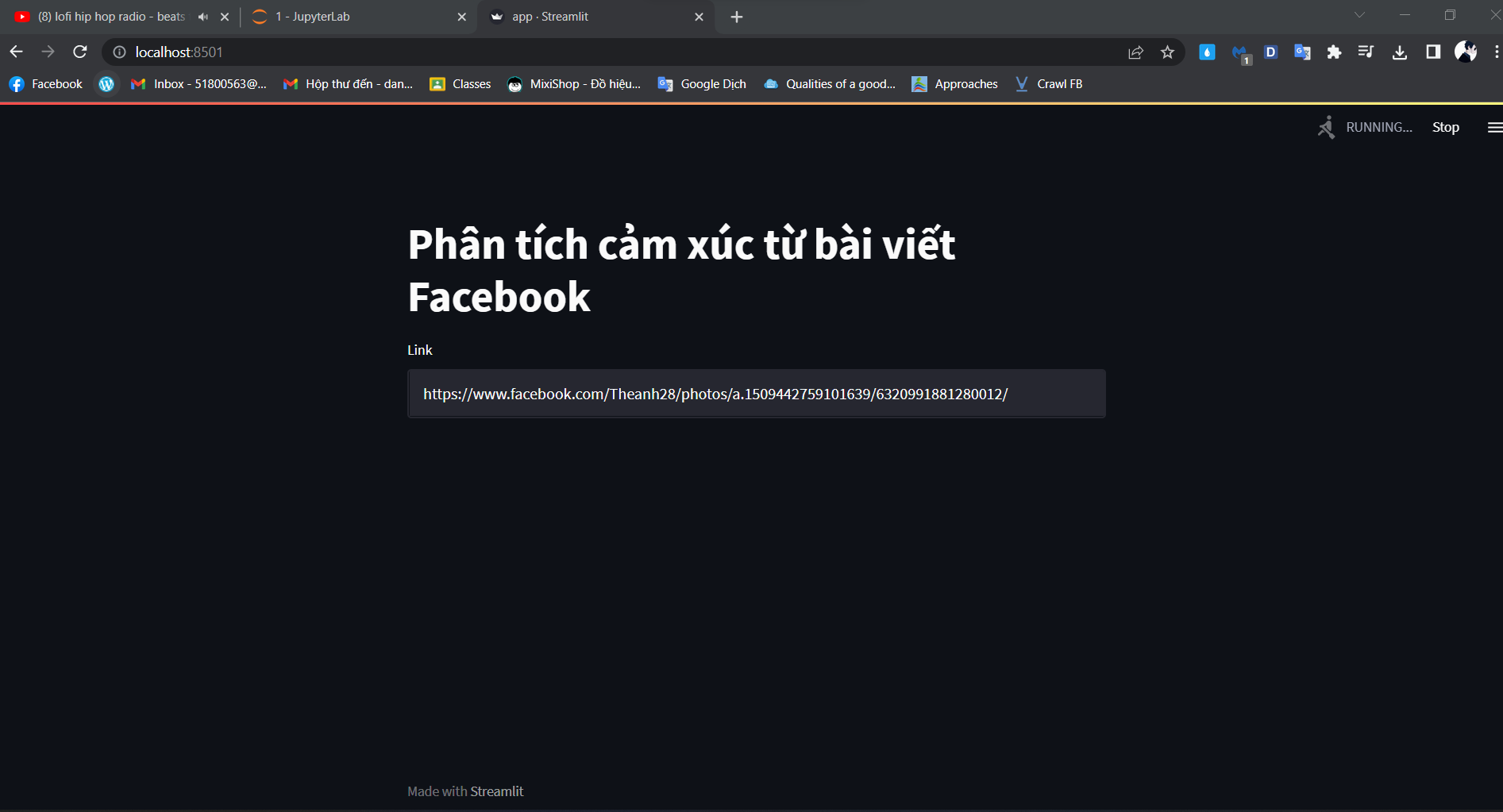
Hình 4.13 Dùng câu lệnh để khởi động web

Sau khi chạy được giao diện như sau:



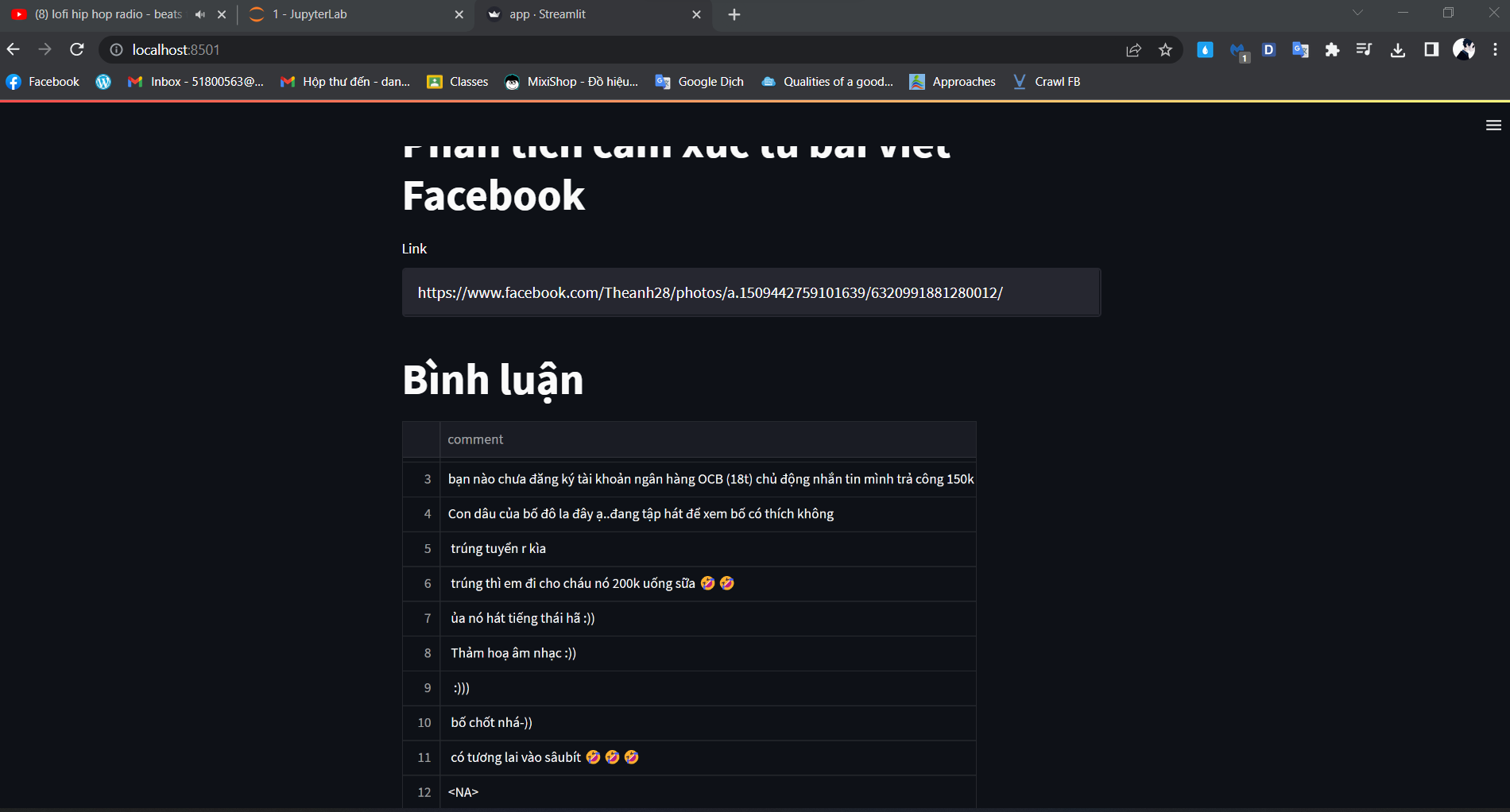
Hình 4.14 Giao diện web phân tích cảm xúc từ bài viết Facebook

Nhập link từ một bài viết bất kỳ vào ô input

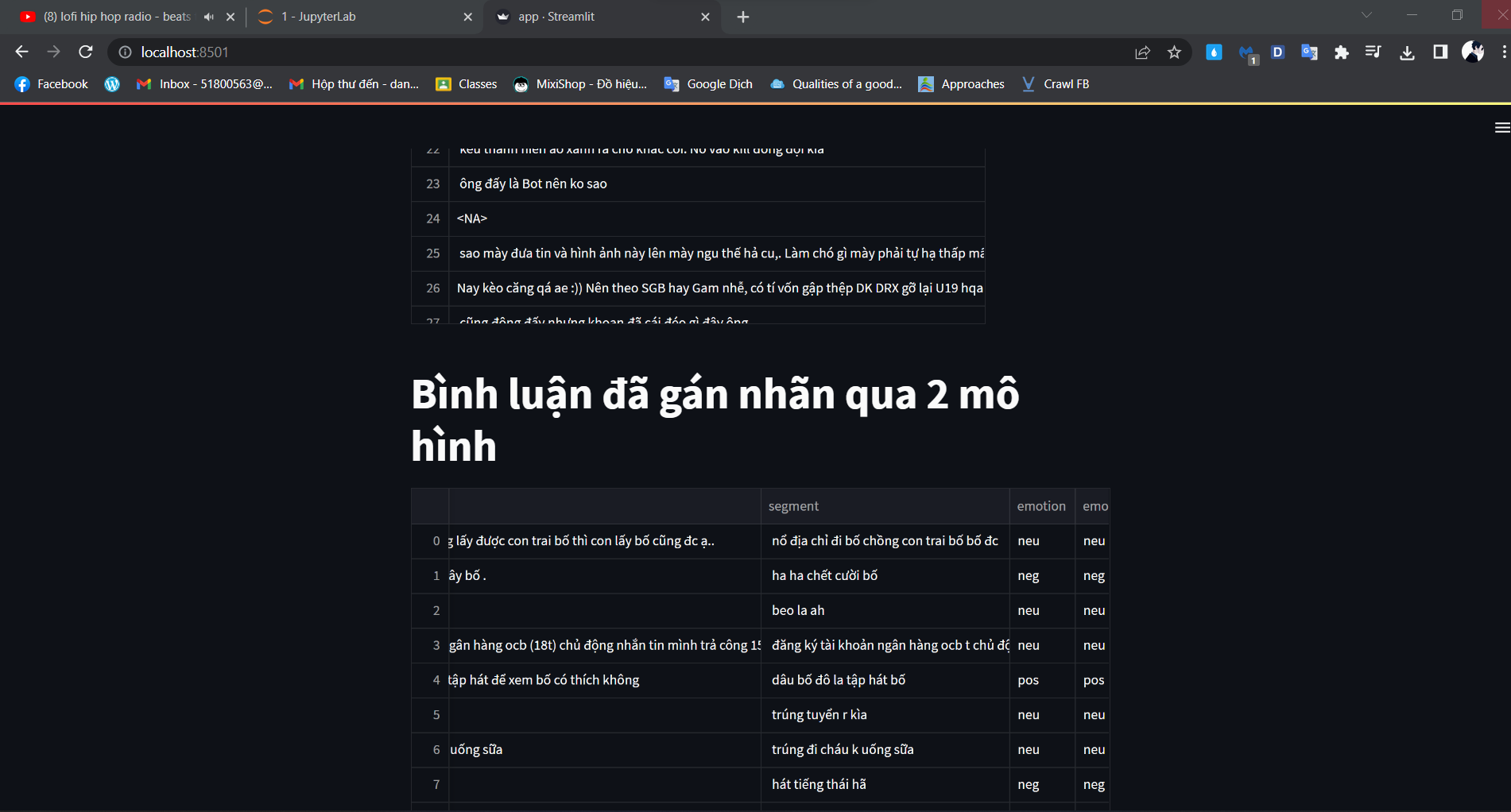


Hình 4.15 Nhập link vào ô input

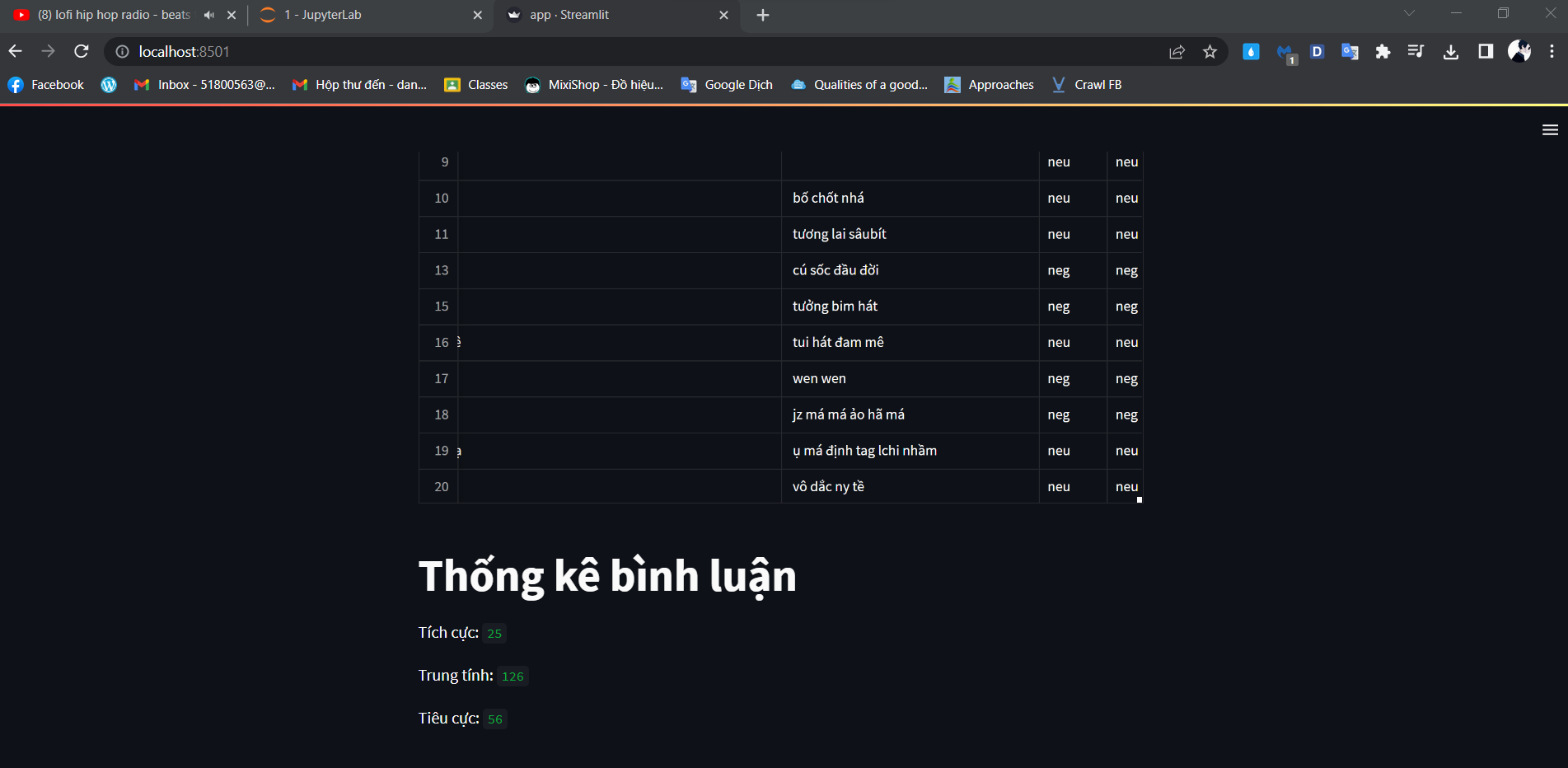
Sau đó nhấn enter, sau vài phút nhận được kết quả như sau:



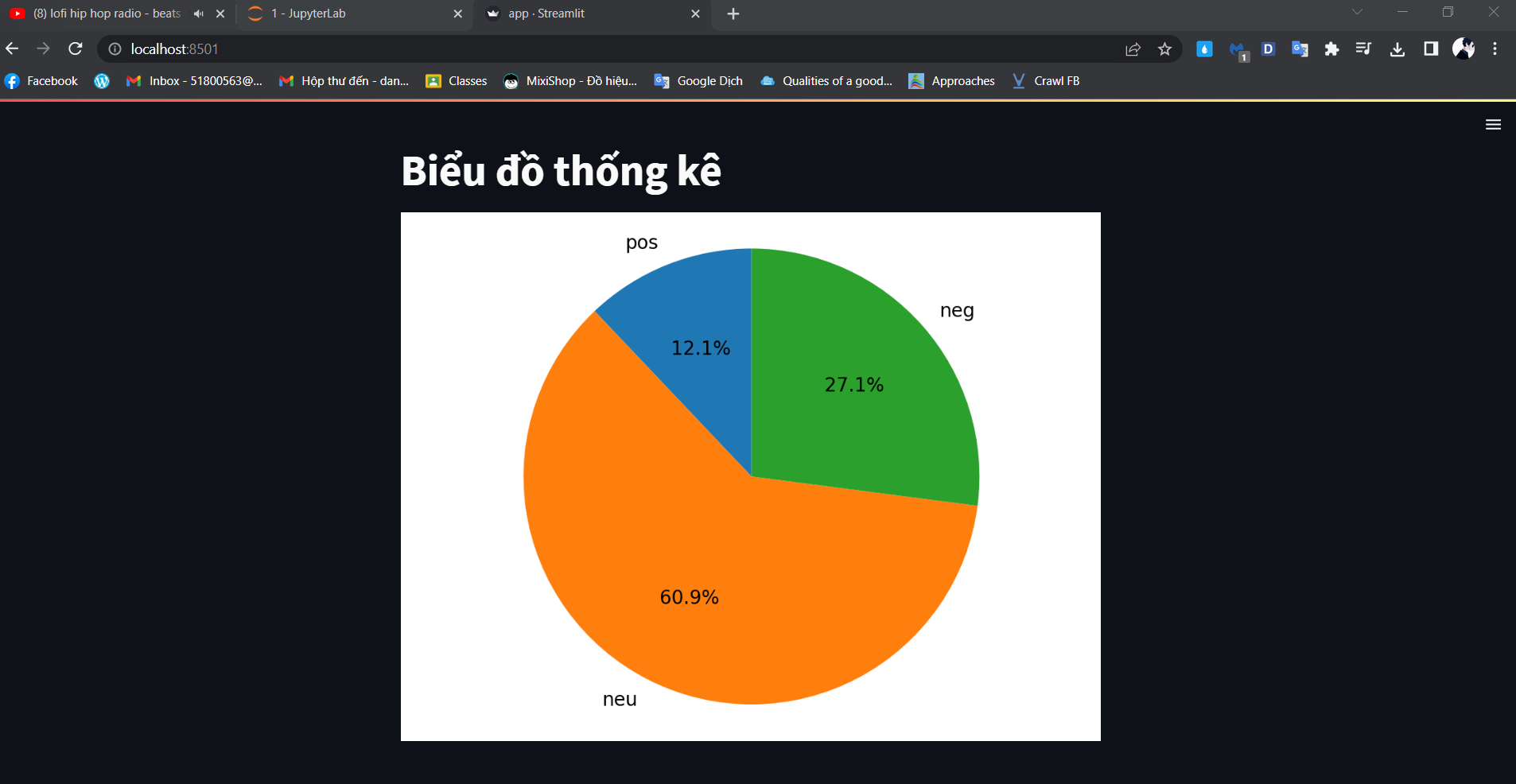
Hình 4.16 Giao diện các bình luận đã được crawl



Hình 4.17 Giao diện các bình luận đã được gán nhãn qua 2 mô hình



Hình 4.18 Giao diện thống kê bình luận qua các nhãn



Hình 4. 19 Giao diện của biểu đồ thống kê các nhãn

**TÀI LIỆU THAM KHẢO**

**Tiếng Việt**

1. https://viblo.asia/p/lam-the-nao-crawl-1-trieu-comments-tren-facebook-Qpmle1NVlrd
2. https://github.com/stopwords/vietnamese-stopword
3. https://machinelearningcoban.com/2017/01/27/logisticregression/?fbclid=IwAR05yt5ZehVrNwX0lMQ-83o2YV3qhMrSZweIs49y50cg8Ebb6w61g4CjjSs

4. <https://viblo.asia/p/logistic-regression-bai-toan-co-ban-trong-machine-learning-924lJ4rzKPM?fbclid=IwAR168e4V78CwuIgJpoLJnQhhS0S6rItjBT_N70gLbj6kqmjljt8CNKpGogo>

5. https://viblo.asia/p/gioi-thieu-ve-support-vector-machine-svm-6J3ZgPVElmB?fbclid=IwAR2EYaH0rR6gFxw176zQTwT\_rPymYI1eEyff-vZ--PMKulG-\_MRToKQVoZU

**Tiếng Anh**

6. https://www.meaningcloud.com/blog/accuracy-in-sentiment analysis?fbclid=IwAR2dXavIhVyaJaNJLqgCj310wg1QAX4MzBYTiGHBWxxZKb7He587zf5DKEs